Nouvelles paramétrisations de réseaux Bayésiens et leur estimation implicite - Famille exponentielle naturelle et mélange infini de Gaussiennes

Aida Jarraya Siala

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Aida JARRAYA SIALA

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
Docteur de la Faculté des Sciences de Sfax
sous le label de l'Université de Nantes Angers Le Mans

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique, section CNU 27
Unité de recherche : Laboratoire d'informatique de Nantes-Atlantique (LINA)

Soutenue le 26 octobre 2013
Thèse n° : ED 503-210

Nouvelles paramétrisations de réseaux Bayésiens et leur estimation implicite

Famille exponentielle naturelle et mélange infini de Gaussiennes

JURY

Président :	**M. Abdelhamid HASSAIRI**, Professeur, Université de Sfax, Tunisie
Rapporteurs :	**M. Rafik AGUECH**, Professeur, Université de Monastir, Tunisie
	M. Richard EMILION, Professeur des Universités, Université d'Orléans
Examinateur :	**M. Jean-Michel POGGI**, Professeur, Université Paris Descartes
Directeurs de thèse :	**M. Philippe LERAY**, Professeur des Universités, Université de Nantes
	M. Afif MASMOUDI, Professeur, Université de Sfax

Table des matières

Introduction

Ce travail s'inscrit dans une thématique de recherche commune au laboratoire de Probabilités et Statistique de l'université de Sfax et à l'équipe Connaissances et Décision du Laboratoire d'Informatique de Nantes Atlantique de l'université de Nantes. Il présente une synthèse des travaux de recherche que nous avons réalisé dans le domaine de probabilités et statistique, plus précisément sur l'apprentissage Bayésien de modèles graphiques probabilistes.
Les réseaux Bayésiens et plus généralement les modèles graphiques probabilistes sont un formalisme de raisonnement probabiliste en plein essor pour la fouille de données et pour la modélisation de systèmes complexes lorsque les situations sont incertaines et/ou les données sont incomplètes [59, 38, 57].
Les réseaux Bayésiens sont des modèles graphiques qui représentent les relations probabilisées entre un ensemble des variables [11, 57, 10]. Ces dernières années, les réseaux Bayésiens sont devenus un outil populaire pour représenter et manipuler des connaissances expertes dans un système expert [16].
Les réseaux Bayésiens sont souvent utilisés car ils ont plusieurs avantages par rapport à d'autres techniques.
Les réseaux Bayésiens peuvent représenter intuitivement un domaine de connaissance, beaucoup d'expériences montrent qu'il est plus facile de formaliser les connaissances sous forme d'un graphe que sous forme d'un système basé sur des règles.
Pour ces raisons, les réseaux Bayésiens sont désormais utilisés dans divers domaines comme la finance, l'économie, la médecine, la robotique, le génie civil, la géologie, la génétique, la criminologie, l'écologie, l'industrie, ...[4, 56].
L'apprentissage d'un réseau Bayésien consiste à estimer le graphe (la structure) G et les paramètres des distributions de probabilités conditionnelles

associées à ce graphe. Cet apprentissage peut s'effectuer aussi bien à partir d'expertises sur le domaine qu'avec des données réelles complètes ou incomplètes [47, 57] et fait l'objet de nombreux travaux de recherche [10, 59, 41].
Les algorithmes d'apprentissage de réseaux Bayésiens [45] utilisent en pratique une approche Bayésienne classique d'estimation a posteriori dont les paramètres sont souvent déterminés par un expert ou définis de manière uniforme [61]. De plus, la recherche du meilleur graphe passe lui aussi par la maximisation d'une loi marginale pour laquelle des hypothèses sont aussi faites concernant la distribution a priori des paramètres (score Bayésien Dirichlet). Le choix de ces paramètres pose souvent des problèmes.
Le coeur de cette thèse concerne l'application aux réseaux Bayésiens de plusieurs avancées dans le domaine des Statistiques comme l'estimation implicite, les familles exponentielles naturelles, ou le mélange infini de lois Gaussiennes pour proposer de nouvelles formes paramétriques, pour l'estimation des paramètres de tels modèles et l'apprentissage de leur structure.
L'estimation implicite proposée par [26] au sein du Laboratoire de Probabilités et Statistique de Sfax est une approche Bayésienne non informative proposée de manière très générale pour de nombreuses familles de distributions de probabilités. Cette approche implicite permet de contourner le problème de choix des a prioris de l'estimation Bayésienne. Des premiers travaux ont montré l'intérêt de cette approche pour l'apprentissage des paramètres des réseaux Bayésiens, dans le cadre de l'estimation des paramètres de lois multinomiales avec des données complètes [4] et incomplètes [5]. Nos objectifs sont donc de poursuivre ces travaux dans le cadre des réseaux Bayésiens dont les variables ne sont pas toutes discrètes car dans la littérature les travaux sont concentrés généralement sur des modèles à variables discrètes alors que le formalisme des réseaux Bayésiens permet de traiter des modèles plus complexes, soient complètement continus, soient mixtes discrets/continus.
Nous avons aussi développé une approche consistant à étendre les paramétrisations classiques des réseaux Bayésiens, en les élargissant aux familles exponentielles naturelles. Nous avons commencé par nous intéresser au cas discret, en proposant une nouvelle paramétrisation appelée réseau Bayésien exponentiel discret. Cette proposition permet de traiter des variables discrètes de domaine infini, avec un nombre de paramètres limités, ce que ne

peuvent pas faire les réseaux Bayésiens discrets classiques.
Cette approche se poursuit ensuite avec l'application des mélanges infini de lois Gaussiennes aux réseaux Bayésiens. Ce travail, permet de proposer une paramétrisation non Gaussienne qui pourrait compléter les modèles de type LiNGAM (Linear, Non-Gaussian, Acyclic causal Models) proposés par Shimizu *et al* [64], qui se concentrent sur l'apprentissage de la structure mais ne disposent pas de forme paramétrique.

Ce rapport est organisé comme suit :
Dans le premier chapitre, nous présentons les différents types de réseaux Bayésiens, leurs définitions et leurs propriétés fondamentales. Pour plus de détails le lecteur pourra se référer à l'ouvrage de Pearl [59] et à celui de Naïm et al [57].
Dans le deuxième chapitre, nous rappelons des généralités sur les familles exponentielles naturelles, puis nous présentons la méthode implicite. Nous présentons également le mélange infini de Gaussiennes.
Dans le troisième chapitre, nous présentons notre première contribution, une nouvelle paramétrisation de réseaux Bayésiens à l'aide de lois discrètes appartenant à des familles exponentielles naturelles et nous proposons aussi les estimateurs Bayésiens de cette nouvelle paramétrisation. Nous nous sommes penchés sur l'apprentissage de la structure et par la suite nous avons développé une nouvelle fonction score qui étend le score Dirichlet Bayésien. Nous avons déterminé de façon empirique dans quels contextes les réseaux Bayésiens exponentiels discrets peuvent être une bonne alternative par rapport aux réseaux Bayésiens usuels pour l'estimation de densité. Les premiers résultats concernant ces modèles ont été publiés dans les actes de la conférence *23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI'2011)* [33]. Les résultats concernant ces modèles pour l'estimation de densité ont été présentés et publiés dans les actes de la conférence *Eighth International Conference on Intelligent Computing (ICIC 2012)* [34]. Une version longue de ces travaux est acceptée pour publication dans la revue Neurocomputing [36].
Dans le quatrième chapitre, nous montrons l'intérêt de l'approche Implicite pour l'apprentissage des paramètres dans un modèle à volatilité stochastique.

Ce résultat particulier a fait l'objet d'une publication dans la revue Communications in Statistics Theory and Methods [25]. Nous avons effectué aussi une étude comparative de l'estimation des paramètres d'un réseau Bayésien Gaussien conditionnel par la méthode implicite et par les méthodes Bayésiennes classiques. Ces derniers résultats ont été publiés dans les actes de la conférence *10th International FLINS Conference on Uncertainty Modeling in Knowledge Engineering and Decision Making (FLINS 2012)* [35]. Cet article a été sélectionné pour une soumission longue à paraître dans la revue IJCIS (International Journal of Computational Intelligence Systems) [37].

Dans le cinquième chapitre, nous introduisons la notion de mélange infini de lois Gaussiennes dans les réseaux Bayésiens et nous étudions les réseaux Bayésiens associés à un modèle mélange infini de lois Gaussiennes. Nous introduisons dans un premier temps la notion de mélange infini de lois Gaussiennes dans les réseaux Bayésiens. Nous passons ensuite à l'étude de l'apprentissage en reliant ces nouveaux modèles aux modèles linéaires non gaussiens. Cette dernière partie de notre travail de thèse n'a pas encore fait l'objet de publication.

Chapitre 1

Réseaux Bayésiens

1.1 Introduction

Les modèles graphiques probabilistes fournissent un outil pour la gestion simultanée de deux problèmes "l'incertain" et la "complexité". Les modèles graphiques bénéficient non seulement des avantages des modèles probabilistes, mais, de plus, ils présentent des avantages supplémentaires liés à leur représentation graphique [39]. En effet, le coté graphique de ces modèles fournit à la fois une interface intuitive par laquelle on peut modéliser un grand nombre de variables aléatoires en représentant les indépendances conditionnelles entre elles ainsi qu'une structure de données. Le coté probabiliste des modèles graphiques permet de factoriser une distribution de la loi jointe qui relie les variables. Le coté graphique permet en plus de visualiser l'architecture du modèle complexe représentant le problème ainsi que les relations entre les différentes parties du modèle.
Ce chapitre est consacré à l'étude des modèles graphiques probabilistes de différents points de vues. Nous commençons par introduire le principe général de ces modèles graphiques en mettant l'accent principalement sur l'indépendance et l'apprentissage. Nous détaillons ensuite la notion d'apprentissage des paramètres et de structure dans une classe particulière des modèles graphiques probabilistes : les Réseaux Bayésiens. Nous étudions différentes déclinaisons des réseaux bayésiens, les RB discrets, les RB gaussiens, les RB gaussiens conditionnels ainsi que des modèles linéaires non gaussiens.

1.2 Modèle graphique

1.2.1 Définition

Les modèles graphiques probabilistes (MGP) allient la théorie des probabilités et la théorie des graphes.
Les modèles graphiques probabilistes se divisent alors en plusieurs classes selon la nature de la structure du graphe.
Si le graphe est un graphe orienté sans circuit le MGP est un réseau Bayésien [59]. Si le graphe est un graphe non orienté le MGP est un champ de Markov [44]. Si le graphe est semi-dirigé (certain arcs sont orientés et d'autres sont non orientés) le MGP est un Chain graph [42].
Nous nous intéressons plus particulièrement aux modèles graphiques dirigés.

1.2.2 Modèles d'indépendance

Les modèles graphiques ont un grand intérêt pour modéliser les relations d'indépendances conditionnelles entre les variables aléatoires. Un modèle graphique de dépendances est une représentation composée d'une paire d'ensembles $G = (V, E)$, où V est un ensemble fini de noeuds et l'ensemble E des arcs est un sous ensemble de $V \times V$ de paires ordonnées de noeuds distincts. Un graphe est alors une structure simple ne contenant ni des arcs multiples ni des cycles. Un tel modèle graphique est alors un modèle de dépendances défini par la définition suivante.

Définition 1.1. *Un graphe de dépendance $G = (V, E)$, orienté ou non, est composé d'un ensemble de noeuds V , représentant des variables aléatoires et un ensemble d'arêtes E tels que l'absence d'une arête entre les noeuds u et v représente une relation d'indépendance conditionnelle entre les variables aléatoires associées à ces deux noeuds.*

Définition 1.2. *Soient A, B et C trois variables aléatoires. A et B sont indépendantes conditionnellement à C si la loi conditionnelle de A, B sachant C est égale au produit des lois de A sachant C et de B sachant C.*

Les modèles graphiques probabilistes représentent graphiquement les indépendances. L'ensemble V de noeuds est en bijection avec l'ensemble des variables aléatoires. Un noeud représente une variable aléatoire. La topologie

de ce graphe est utilisée pour représenter les dépendances entre les variables. Ainsi, un arc entre deux noeuds indique une relation de dépendance entre les deux variables associées et l'absence d'un arc indique une relation d'indépendance. Un modèle graphique probabiliste est capable de représenter des indépendances conditionnelles au sein d'un groupe de variables (voir définition 1.2) grâce à la notion de séparation, ceci se traduit par les propriétés de Markov au sein du graphe lui même. La lecture des indépendances conditionnelles sur un graphe est liée à la notion de séparation. La séparation est un critère permettant de statuer si deux sous ensembles de sommets disjoints d'un graphe sont ou non séparés l'un de l'autre étant donné un troisième sous ensemble disjoint. La séparation est définie différemment selon le type de graphe auquel on s'intéresse (orienté ou non-orienté notamment). Ici, nous nous limitons à la définition de la séparation dans le cadre des graphes orientés, appelée d-séparation.

Définition 1.3. *On dit que X et Y sont d-séparés par Z si pour tous les chemins entre X et Y l'une au moins des deux conditions suivantes est vérifiée :*

• *Le chemin converge en un noeud W, tel que $W \neq Z$ et W n'est pas une cause directe de Z.*

• *Le chemin passe par Z, est soit divergent soit en série au noeud Z.*

Si X et Y sont d-séparés par Z ($i.e; X \perp_d Y/Z$), alors X et Y sont indépendantes conditionnellement à Z. Ce résultat est très important, car il permet de limiter les calculs de probabilités grâce à des propriétés du graphe.

Définition 1.4. *Deux DAG G_1 et G_2 sont équivalents s'ils impliquent les mêmes indépendances conditionnelles.*

Verma et Pearl [66] ont montré que deux graphes sont équivalents si et seulement si ces deux graphes ont le même squelette et les mêmes V-structures (sous-graphes du type $A \longrightarrow B \longleftarrow C$ où A et C ne sont pas directement reliés).

Définition 1.5. *Soit P la distribution de la probabilité jointe d'un ensemble de variables aléatoires $V = (X_1, X_2, \ldots, X_n)$ et soit $G = (V, E)$ un graphe dirigé acyclique. On dit que le couple (G, P) vérifie :*

La propriété orientée de Markov locale *(OML) si conditionnellement à ses parents, toute variable X_i est indépendante de l'ensemble de ses non-descendants ($ndesc(X_i)$).*

$$i.e, \quad \forall X_i \in V, X_i \perp ndesc(X_i)/pa(X_i).$$

avec $pa(X_i)$ est l'ensemble des parents d'un noeud X_i.

Plus formellement un réseau Bayésien (RB) se définit de la manière suivante.

Définition 1.6. *Soit P la loi de la probabilité jointe d'un vecteur aléatoire $(X_1, X_2, \ldots, X_n)$ et soit G un graphe dirigé acyclique où l'ensemble de ses noeuds sont notés aussi $X_1, X_2, \ldots, X_n$ (à chaque variable aléatoire X_i est associé un noeud noté aussi X_i). On dit que (G, P) est un réseau Bayésien s'il vérifie la propriété orientée de Markov locale.*

Théorème 1.1. *Soit un graphe dirigé acyclique G tel que à chaque noeud X_i, on associe une variable aléatoire notée aussi X_i et une loi de probabilité de chaque X_i sachant la configuration de ses parents $pa(X_i)$ dans G. Alors le produit de ces probabilités conditionnelles donne une probabilité jointe de $(X_1, \ldots, X_n)$*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i/pa(X_i))$$

et (G, P) vérifie la propriété orientée de Markov locale.

Cette décomposition d'une fonction globale en un produit de termes locaux dépendant uniquement du noeud considéré et de ses parents dans le graphe est une propriété fondamentale des réseaux Bayésiens. Elle est à la base des premiers travaux portant sur le développement d'algorithmes d'inférence, qui calculent la probabilité de n'importe quelle variable du modèle à partir de l'observation même partielle des autres variables. Ce problème a été prouvé NP-complet, mais a abouti à différents algorithmes qui peuvent être assimilés à des méthodes de propagation d'information dans un graphe. Ces méthodes utilisent évidemment la notion de probabilité conditionnelle.
Nos travaux ne concernent pas ces algorithmes d'inférence, par contre on se concentre sur la partie d'apprentissage.

Il faut noter que l'appellation "réseaux Bayésiens" prête à confusion. En effet, ceux-ci ne sont pas forcément des modèles Bayésiens, au sens statistique du terme. Ce sont des modèles graphiques probabilistes utilisant le théorème de Bayes pour "raisonner".
Nous allons présenter différentes déclinaisons des réseaux bayésiens, les RB discrets (section 1.3), les RB gaussiens (section 1.4), les RB gaussiens conditionnels (section 1.5) ainsi que des modèles linéaires non gaussiens proches des précédents modèles (section 1.6).

1.2.3 Apprentissage

Un réseau Bayésien est constitué à la fois d'un graphe et d'un ensemble de probabilités conditionnelles.
L'apprentissage d'un réseau Bayésien doit donc répondre aux deux questions suivantes :
• Comment estimer les lois de probabilités conditionnelles ?
• Comment trouver la structure du réseau Bayésien ?
Nous allons donc séparer le problème de l'apprentissage en deux parties :
• L'apprentissage des paramètres, où nous supposerons que la structure du réseau a été fixée, et où il faudra estimer les probabilités conditionnelles de chaque noeud du réseau.
• L'apprentissage de la structure, où le but est de trouver le meilleur graphe représentant la tâche à résoudre.
En faisant l'hypothèse classique d'indépendance des paramètres, les paramètres des distributions de probabilités conditionnelles de chaque variable sont estimés indépendamment des autres variables.
Les méthodes d'apprentissage des réseaux Bayésiens utilisent en pratique la méthode du maximum de vraisemblance et fréquemment l'approche Bayésienne pour estimer les paramètres du modèle. Dans la méthode Bayésienne, les lois a prioris des paramètres sont souvent déterminées par un expert ou définies de manière uniforme. De plus, les algorithmes d'apprentissage de la structure c'est-à-dire la recherche d'un "bon" graphe, qui correspond aux données et aux a prioris, se base aussi sur la maximisation d'une fonction, appelée fonction score, sur laquelle des hypothèses sont aussi faites concer-

nant la distribution a priori des paramètres.
Généralement, l'apprentissage des paramètres se fait en utilisant la méthode du maximum de vraisemblance. Si cette approche paraît naturelle, elle présente néanmoins un inconvénient majeur dans le cas d'une base d'apprentissage de taille limitée ; si une configuration particulière peut exister, avec une probabilité faible non nulle et qu'elle n'est pas présente dans la base de données, alors l'approche du maximum de vraisemblance aboutit à une estimation de la probabilité nulle. Or, le fait qu'une configuration n'a pas été observée ne signifie pas nécessairement qu'elle a une probabilité nulle. Afin d'y remédier, il serait bon de pouvoir exprimer et quantifier la possibilité de la survenance d'un tel événement. C'est cet objectif que remplit l'approche Bayésienne (voir Annexe B).

1.3 Réseaux Bayésiens discrets

1.3.1 Définition

Un réseau Bayésien dont toutes les variables sont discrètes est appelé un réseau Bayésien discret. Considérons un réseau Bayésien formé de n variables aléatoires $X = (X_1, \ldots, X_n)$ et d'un graphe G dirigé acyclique (DAG).
On suppose que chaque variable aléatoire X_i passe par un nombre fini d'états r_i et ses parents $pa(X_i)$ possèdent q_i configurations, c'est à dire $q_i = \prod_{X_j \in pa(X_i)} r_j$.
$\theta_i = P(X_i/pa(X_i))$ est la matrice des probabilités conditionnelles du noeud i connaissant l'état de ses parents $pa(X_i)$. Un paramètre θ_i est un tableau contenant l'ensemble des probabilités de la variable X_i pour chacune de ses valeurs possibles sachant chacune des valeurs prises par l'ensemble de ses parents $pa(X_i)$. On désigne par $\theta_{i,j,k}$ la probabilité $P(X_i = k/pa(X_i) = j)$.
Un exemple de réseau bayésien discret est donné dans la figure 1.1. Il s'agit d'un réseau décrivant les relations conditionnelles existant entre :
- la survenue éventuelle d'un séisme,
- la diffusion d'un flash radio (ou TV) annonçant un séisme,
- le cambriolage d'un édifice,
- le déclenchement de l'alarme de cet édifice, suite à un séisme ou un cambriolage,

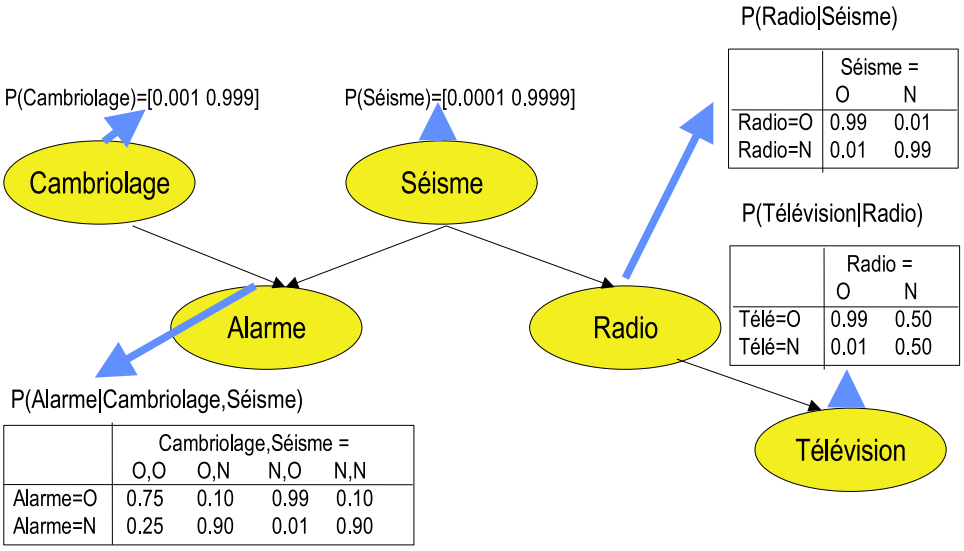

FIG. 1.1 – Exemple de réseau Bayésien discret

A chaque sommet du graphe est associée une table de probabilités permettant de déterminer la probabilité avec laquelle la variable associée peut prendre une valeur particulière étant donné celles prises par ses parents (s'ils existent).

1.3.2 Apprentissage des paramètres

On désigne par D la base de données et $N_{i,j,k}$ le nombre de fois d'observer X_i à l'état k et ses parents à la configuration j.
$(N_{ij1}, \ldots, N_{ijr_i})$ est une multinomiale.

Apprentissage Statistique

Dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée est l'estimation du maximum de vraisemblance (MV) qui consiste à maximiser $P(D/\theta)$

$$\widehat{P}(X_i = k/pa(X_i) = j) = \widehat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_{k=1}^{r_i} N_{i,j,k}}$$

Apprentissage Bayésien

L'utilisation du théorème de Bayes permet de proposer une formulation Bayésienne de l'apprentissage. Il nous paraît important de souligner que l'apprentissage Bayésien n'est pas lié de façon exclusive aux réseaux de même nom. On peut estimer ceux-ci avec d'autres méthodes par exemple par des méthodes statistiques comme on l'a montré ci-dessus.

L'estimation Bayésienne suit un principe différent. Il consiste à trouver les paramètres θ les plus probables sachant que les données ont été observées, en utilisant des a prioris sur les paramètres. La formule de Bayes nous énonce que la loi de probabilité de θ sachant les données D est proportionnelle au produit de la loi de D sachant θ et la loi de θ :

$$P(\theta/D) \propto P(D/\theta)P(\theta).$$

Lorsque la distribution de l'échantillon suit une loi multinomiale

$$P(D/\theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{i,j,k}^{N_{i,j,k}}$$

où $\theta = (\theta_{i,j,k})_{1 \leq i \leq n, 1 \leq j \leq q_i, 1 \leq k \leq r_i}$.

La loi a priori conjuguée est la distribution de Dirichlet.

On suppose que $\theta_{ij} = (\theta_{i,j,1}, \ldots, \theta_{i,j,r_i}) \sim Dir(\alpha_{i,j,1}, \ldots, \alpha_{i,j,r_i})$. Ainsi

$$P(\theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{i,j,k})^{\alpha_{i,j,k}-1}$$

où $\alpha_{i,j,k}$ sont les coefficients de la distribution de Dirichlet associée à la loi a priori.

La distribution a posteriori des paramètres est alors définie par

$$P(\theta/D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{i,j,k})^{N_{i,j,k}+\alpha_{i,j,k}-1}.$$

L'approche du maximum a posteriori (MAP) consiste à rechercher le jeu de paramètres θ dont la probabilité a posteriori, c'est-à-dire sachant que la base de données D a été observée, est maximale.

L'estimateur $\widehat{\theta}_{ijk}^{MAP}$ Bayésien MAP est défini par

$$\widehat{P}(X_i = k/pa(X_i) = j) = \widehat{\theta}_{i,j,k}^{MAP} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_{k=1}^{r_i} (N_{i,j,k} + \alpha_{i,j,k} - 1)}$$

où $\alpha_{i,j,k}$ sont les paramètres de la distribution de Dirichlet associée à la loi a priori.
Une autre appproche Bayésienne consiste à calculer l'espérance a posteriori des paramètres $\theta_{i,j,k}$ au lieu d'en chercher le maximum. L'estimateur d'espérance a posteriori (EAP) de θ_{ijk} est définie par

$$\widehat{P}(X_i = k/pa(X_i) = j) = \widehat{\theta}^{EAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k}}{\sum_{k=1}^{r_i}(N_{i,j,k} + \alpha_{i,j,k})}.$$

1.3.3 Apprentissage de la structure

Nous avons examiné différentes méthodes d'apprentissage des paramètres d'un réseau Bayésien à partir des données complètes en supposant que la structure de ce réseau était déjà connue. Se pose maintenant le problème de l'apprentissage de cette structure : comment trouver le meilleur graphe qui représente le réseau Bayésien [57, 45, 10, 18].
Une première approche consiste à retrouver les indépendances conditionnelles entre les variables. Tandis qu'une deuxième essaie de quantifier l'adéquation d'un réseau Bayésien au problème à résoudre, c'est-à-dire d'associer un score à chaque réseau Bayésien. Puis elle recherche la structure qui donnera le meilleur score dans l'espace des graphes orientés sans circuit. Une troisième famille d'algorithmes a vu le jour pour apprendre la structure optimale d'un RB : les méthodes hybrides qui permettent de combiner l'apprentissage par recherche d'indépendances conditionnelles et celui à base de scores afin de profiter des avantages de chacune des deux approches.
Une approche exhaustive est impossible en pratique en raison de la taille de l'espace de recherche. Robinson [62] a prouvé que le nombre de structures possibles à partir de n noeuds est super-exponentiel. Dans notre travail, on se concentre sur les approches basées sur une fonction score. Ces approches vont soit chercher la structure qui maximise un certain score, soit chercher les meilleures structures et combiner leurs résultats [45].

Propriétés importantes des fonctions de score

Nous rappelons maintenant des propriétés importantes en pratique concernant les fonctions scores.

Définition 1.7. *Un score S est dit décomposable s'il peut être écrit comme une somme ou un produit de scores locaux dont chacune n'est fonction que seulement d'un noeud et de l'ensemble de ses parents. Si n est le nombre de noeuds du graphe G, le score doit s'écrire sous l'une des formes suivantes :*

$$S(G) = \sum_{i=1}^{n} s(X_i, pa(X_i)) \text{ ou } S(G) = \prod_{i=1}^{n} s(X_i, pa(X_i))$$

Pour que les approches à base de score soient réalisables en pratique, nous verrons que le score doit être décomposable localement, car posséder un score calculable localement, permet de n'estimer que la variation de ce score entre deux structures voisines, au lieu de le recalculer entièrement pour chaque nouvelle structure.

Définition 1.8. *Un score qui associe une même valeur à deux graphes équivalents est dit équivalent.*

Nous avons vu dans la définition 1.4 que certains réseaux bayésiens différents mais avec le même squelette permettaient de modéliser la même décomposition de la loi jointe de l'ensemble des variables. Il serait alors intéressant d'associer une même valeur du score à toutes les structures équivalentes.

Quelques fonctions de score

En général les scores existants sont tous des approximations différentes de la vraisemblance marginale [11].
Plusieurs fonctions de score peuvent être utilisées, nous citons le score BD (Bayesian Dirichlet), le score BDe (Bayesian Dirichlet Equivalent), le score BDeu, le score BD_γ (Bayesian Dirichlet généralisé), l'entropie conditionnelle, le score AIC (Akaike Information Criterion), le score BIC et la longueur de description minimale (MDL) (voir [57] pour plus de détails sur ces scores). Comme nous nous intéressons dans ces travaux à l'apprentissage bayésien, nous allons décrire plus particulièrement le score bayésien proposé en premier dans la littérature BD, et sa "correction" BDe, plus couramment utilisée.

Le score BD (Dirichlet Bayésien)

Cooper et Herskovits [15] proposent un score basé sur une approche Bayésienne. En partant d'une loi a priori sur les structures possibles $P(G)$, le but est d'exprimer la probabilité a posteriori des structures possibles sachant que les données D ont été observées $P(G/D)$ ou plus simplement $P(G, D)$:

$$\begin{aligned} ScoreBD(G,D) = \quad P(G,D) \quad &= \int_{\Theta} L(D/\theta, G) P(\theta/G) P(G) d\theta \\ &= P(G) \int_{\Theta} L(D/\theta, G) P(\theta/G) d\theta \end{aligned}$$

où $L(D/\theta, G)$ désigne la fonction de vraisemblance.
Avec les hypothèses classiques d'indépendance des exemples, et en prenant une distribution a priori de Dirichlet sur les paramètres il est néanmoins possible d'exprimer le score BD facilement

$$ScoreBD(G,D) = P(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \qquad (1.1)$$

où Γ est la fonction Gamma.
Il est à noter que le score BD est décomposable.

Le score BDe (Dirichlet Bayésien équivalent)

Ce critère suggéré par Heckerman [29] s'appuie sur la même formule que le score Dirichlet Bayésien avec des propriétés supplémentaires intéressantes, comme la conservation du score pour des structures équivalentes. Le score BDe utilise une distribution a priori sur les paramètres définie par :

$$\alpha_{ijk} = N' P(X_i = k, pa(X_i) = j / G_c) \qquad (1.2)$$

où G_c est la structure a priori n'encodant aucune indépendance conditionnelle (graphe complètement connecté) et N' est un nombre d'exemples "équivalent" défini par l'utilisateur.
Heckerman *et al* [10] prouvent aussi que le score BDe utilisant les a priori définis par l'équation (1.2) n'a plus besoin d'utiliser une distribution de Dirichlet comme loi a priori sur les paramètres. Heckerman et Geiger [28] montrent que le score BDe est équivalent.
Etant donné une base de données D, un nombre d'exemples équivalent défini

par l'utilisateur, une distribution de probabilité jointe $P(D/G_c)$, on considère la fonction suivante

$$l(X) = \prod_k \frac{\Gamma(N'P(X = k/G_c) + N_k)}{\Gamma(N'P(X = k/G_c))}$$

où k est un état de X, et N_k est le nombre de cas dans D dans laquelle $X = k$. Ensuite le terme de vraisemblance du BDe devient

$$P(D/G) = \prod_{i=1}^{n} \frac{l(\{x_i\} \cup pa(X_i))}{l(pa(X_i))}. \tag{1.3}$$

Une structure peut être transformée en une structure équivalente par une série d'inversion d'arc. On démontre que le score BDe satisfait l'équivalence en général, si on peut le faire pour le cas où deux structures équivalentes différent par une seule inversion d'arc. Soit G_{s_1} et G_{s_2} deux structures équivalentes qui différent seulement dans la direction de l'arc entre x_i et x_j ($x_i \longrightarrow x_j$ dans G_{s_1}).

Soit R l'ensemble des parents de x_i dans G_{s_1}. $R \cup \{x_i\}$ est l'ensemble des parents de x_j dans G_{s_1}, R est l'ensemble des parents de x_j dans G_{s_2} et $R \cup \{x_j\}$ est l'ensemble des parents de x_i dans G_{s_2}. Puisque deux structures différent uniquement par l'inversion d'un seul arc, les seuls termes dans le produit de l'équation (1.3), qui peuvent différer sont celles qui concernent x_i et x_j. Pour G_{s_1}, ces termes sont

$$\frac{l(\{x_i\} \cup R)}{l(R)} \frac{l(\{x_i, x_j\} \cup R)}{l(\{x_i\} \cup R)}$$

tandis que pour G_{s_2}

$$\frac{l(\{x_j\} \cup R)}{l(R)} \frac{l(\{x_i, x_j\} \cup R)}{l(\{x_j\} \cup R)}.$$

Ces termes sont égaux, et donc $P(D/G_{s_1}) = P(D/G_{s_2})$.

Heuristiques de recherche

Se pose aussi le problème de parcours de l'espace des réseaux Bayésiens à la recherche de la meilleure structure. Comme la recherche exhaustive n'est pas une méthode réalisable en pratique lorsque le nombre de noeuds est élevé, il faut utiliser des heuristiques de recherche.

Une première méthode consiste à remplacer l'espace de recherche (espace des DAG) par un espace plus petit, l'espace des arbres. Cette idée consiste à rechercher l'arbre optimal, l'arbre de recouvrement maximal (MWST ou Maximal Weight Spanning Tree [13]) : c'est l'arbre qui relie tous les noeuds en maximisant un critère (un score défini pour tous les arcs possibles entre les variables). Une deuxième idée consiste à ordonner les noeuds pour limiter la recherche des parents possibles pour chaque noeud (algorithme K2 [15]). Une troisième consiste à faire une recherche gloutonne dans l'espace des réseaux bayésiens (algorithme GS [12]). Elle prend un graphe de départ, définit un voisinage de ce graphe, puis associe un score à chaque graphe du voisinage. Le meilleur graphe est alors choisi comme point de départ de l'itération suivante.
Ces méthodes utilisés pour les RBs discret pourront se généraliser à d'autres modèles si l'on a une fonction de score adéquate.

1.4 Réseaux Bayésiens Gaussiens

1.4.1 Définition

Un réseau Bayésien est considéré comme un réseau Bayésien Gaussien (RBG) si la loi de probabilité jointe (JPD) associé à ses variables $Y = (Y_1, \ldots, Y_n)$ est une distribution normale multivariée, $N(\mu, \Sigma)$, de densité

$$f(y) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} exp\{-\frac{1}{2}\ {}^t(y-\mu)T(y-\mu)\}$$

où μ est le vecteur moyen de Y, Σ est la matrice de covariance de taille $n \times n$, avec $|\Sigma|$ est le déterminant de Σ, T est la matrice de précision qui est l'inverse de la matrice de covariance ($T = \Sigma^{-1}$) et on note par ${}^t\mu$ le transposé de μ.
Les réseaux Bayésiens Gaussiens ont été traités, notamment par Neapolitan [58].
Dans un réseau Bayésien Gaussien, la fonction de densité de la distribution jointe peut être exprimée comme le produit de n densités normales unidimensionnelle définies comme

$$f_{Y_i/X_i} \sim N(\mu_i + \sum_{j \in \Lambda_i} b_{ij} y_j, \sigma^2_{i/X_i}) \tag{1.4}$$

avec $X_i = pa(Y_i)$ est l'ensemble des parents continus de Y_i, b_{ij} est le coefficient de régression de Y_j et $\Lambda_i = \{l \in \{1, \ldots, n\}; Y_l \text{ est un parent de } Y_i\}$ et $\sigma^2_{i/X_i} = \Sigma_i - \Sigma_{iX_i}\Sigma^{-1}_{X_i}\ {}^t\Sigma_{iX_i}$ est la variance conditionnelle de Y_i, sachant X_i, où Σ_i est la variance inconditionnelle de Y_i, Σ_{iX_i} est la matrice de covariance entre Y_i et les variables X_i, et Σ_{X_i} est la matrice de covariance de X_i.

1.4.2 Apprentissage des paramètres

On peut estimer les paramètres localement, ou globalement en utilisant le fait qu'un RBG est une gausienne multivariée.

Approche locale

De même que pour le cas discret, grâce à l'indépendance des paramètres, on peut estimer séparément les paramètres de chaque loi conditionnelle de chaque variable. Koller et Friedman [39] détaille comment le principe du MV peut être appliqué dans le cas des réseaux Bayésiens Gaussiens.

On considère le problème de trouver les estimateurs des paramètres d'une variable Gaussienne Y avec des parents continus X, i.e

$$P(y/x) = c|\Sigma|^{-\frac{1}{2}} exp(-\frac{1}{2}(y - Bx - \mu)'\Sigma^{-1}(y - Bx - \mu))$$

où x et y sont des réalisations des variables aléatoires X et Y respectivement, c est une constante.

Considérons un N-échantillon aléatoire. Les statistiques suffisantes sont introduite pour simplifier le calcul des estimateurs :

$$S_{XX'} = \sum_{t=1}^{N} E_t(XX'), S_X = \sum_{t=1}^{N} E_t(X)$$

En appliquant la méthode de maximum de vraisemblance, on obtient les estimateurs suivants :

$$\widehat{\mu}^{MV} = \frac{S_Y - B - S_X}{N}$$

$$\widehat{B}^{MV} = (S_{YX'} - \mu S_{X'})(S_{XX'})^{-1}$$

Pour l'estimation de Σ, on désigne par $B2 = [B\ \mu]$ et $Z = [X\ 1]$, on obtient donc

$$\widehat{\Sigma}^{MV} = \frac{S_{YY'} - S_{YZ'}B2' - B2S_{ZY'} + B2S_{ZZ'}B2'}{N}$$

Approche globale

Il existe une autre méthode d'apprentissage de paramètres, il est aussi possible de transformer un réseau Bayésien Gaussien (RBG) en une distribution Normale multivariée (voir [58]). Il suffit d'effectuer cette transformation afin d'apprendre globalement les paramètres du réseau Bayésien Gaussien. Un tel algorithme d'apprentissage des paramètres d'un RBG est décrit dans le chapitre 7 de [58].

1.4.3 Apprentissage de la structure

Comme pour l'apprentissage des paramètres, l'apprentissage de la structure d'un RBG peut se faire localement (en adaptant les algorithmes existants pour les RB) ou globalement.

Approche locale

La première méthode d'apprentissage à partir de données peut se faire d'une façon analogue au cas discret. On considère le modèle comme un RB et on adapte les algorithmes d'apprentissage de RB. Neapolitan [58] et Geiger et Heckerman [22] discutent de l'apprentissage de structure de réseaux Bayésiens Gaussiens et proposent une déclinaison du score Bayesian Dirichlet équivalent au cas gaussien avec la fonction suivante :

$$P(D/G) = \prod_{i=1}^{n} \frac{P\Big(D_{\{X_i^{(G)} \cup \{y_i\}\}}/G_c\Big)}{P\Big(D_{X_i^{(G)}}/G_c\Big)},$$

où $X_i^{(G)}$ est l'ensemble des parents de Y_i dans G et G_c est le graphe complet.

Approche globale

La deuxième méthode consiste à considérer le modèle comme un modèle de régression complexe dont on cherche à estimer les coefficients non nuls.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\begin{pmatrix} B_{11} & B_{12} & \dots & B_{1n} \\ B_{21} & B_{22} & \dots & B_{2n} \\ \vdots & \vdots & \dots & \vdots \\ B_{n1} & B_{n2} & \dots & B_{nn} \end{pmatrix}}_{A} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} + \begin{pmatrix} B_{01} \\ B_{02} \\ \vdots \\ B_{0n} \end{pmatrix}$$

La matrice A contient des 0 lorsque Y_j ne sont pas des parents de Y_i, sinon elle contient les coefficients de régression.
L'identification de la structure du RBG revient donc à estimer la matrice A, mais en contraignant le plus de termes à être à 0. Pour cela, il est possible de minimiser un terme de pénalisation.
Il existe beaucoup de travaux qui citent des méthodes permettant cette optimisation. Schmidt *et al* [63] montrent que la régression L1 est une technique utile pour l'apprentissage de la structure, que ce soit pour l'accélération de la recherche de l'ordre ou pour l'élagage de l'ensemble des arrêtes possibles. Huang *et al* [31] proposent un algorithme d'apprentissage de structure pour les RBs, réseaux Bayésiens sparse pour l'apprentissage des structures de RBs à partir de données de grandes dimensions.
Huang *et al* [31] ont adopté une nouvelle formulation qui implique un terme de pénalité L1 pour imposer la sparcité pour l'apprentissage et une autre pénalité pour assurer que le réseau Bayésien appris soit un DAG. Ils étudient la propriété théorique pour identifier une valeur finie pour le paramètre de régularisation de la deuxième pénalité, cette valeur assure que le RB appris soit un DAG.

1.5 Réseaux Bayésiens Gaussiens Conditionnels

1.5.1 Définition

Un réseau Bayésien qui contient à la fois des variables discrètes et d'autres absolument continues est dit Gaussien Conditionnel (GC) [42, 40]. Dans ce type de modèles, pour chaque instanciation i d'un ensemble de noeuds discrets Y la distribution conditionnelle des noeuds continus X sachant $Y = i$ est une Normale $N(\mu(i), \Sigma(i))$. Il est à noter que dans le type de modèles que nous considérons dans ce paragraphe, un noeud discret ne peut pas avoir de parents absolument continus. La probabilité conditionnelle d'un noeud discret X_i sachant ses parents $pa(X_i)$ (qui ne peuvent être que discrets) est représentée par une table comme nous l'avons déjà vu pour le cas des réseaux Bayésiens à variables discrètes.
La distribution conditionnelle d'un noeud continu X_i sachant ses parents

$pa(X_i)$ peut être spécifiée par une fonction Gaussienne dont la moyenne est une fonction linéaire des parents et la covariance est fixée (voir [53] pour plus de détails). C'est le modèle de régression linéaire standard.

1.5.2 Apprentissage des paramètres

Dans cette section, nous étudions comment trouver les estimateurs du maximum de vraisemblance (MLEs) des paramètres associés à chaque noeud. Nous avons trois cas. Commençons par le premier cas, si un noeud est discret et ses parents sont discrets on tombe dans l'estimation des paramètres d'un RB discret (section 1.3.2).
Le deuxième cas, si un noeud est continu et ses parents sont tous continus, on est dans le cas de l'estimation des paramètres d'un RBG (section 1.4.2). On passe au troisième cas, où un noeud est continu et ses parents sont continus et discrets. Pour simplifier les équations qui suivent, on utilise les mêmes notations proposés par Murphy [55]. On considère le problème de trouver les estimateurs des paramètres d'une variable Gaussienne conditionnelle Y avec des parents continus X et des parents discrets Q, *i.e*

$$p(y|x, Q = i) = c|\Sigma_i|^{-\frac{1}{2}} exp(-\frac{1}{2}(y - B_i x - \mu_i)'\Sigma_i^{-1}(y - B_i x - \mu_i))$$

où x et y sont des réalisations des variables aléatoires X et Y respectivement, c est une constante et $|y| = d$. Considérons un N-échantillon aléatoire, la fonction vraisemblance est définie par

$$\prod_{t=1}^{N}\prod_{i=1}^{|Q|} p(y_t|x_t, Q_t = i, e_t)^{q_t^i}$$

où $q_t^i = 1$ si Q a la valeur i dans la t^{eme} configuration et 0 sinon. Lorsque Q, X et Y ne sont pas observées, on définit la moyenne de la vraisemblance. Soit $\widetilde{p}(y|x, Q = i) = \exp(l)$ avec

$$l = -\frac{1}{2}\sum_{t=1}^{N} E\Big(\sum_{i=1}^{|Q|} q_t^i \log|\Sigma_i| + q_t^i(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)|e_t\Big).$$

On suppose que

$$E(q_t^i x_t x_t'|e_t) = E(q_t^i|e_t)E(x_t x_t'|Q_t = i, e_t) = w_t^i E_{ti}(XX')$$

où les poids $w_t^i = p(Q = i/e_t)$ sont les probabilités a posteriori et $E_{ti}(XX')$ est le deuxième moment conditionnel. On obtient

$$l = -\frac{1}{2}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i \log|\Sigma_i| - \frac{1}{2}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i E_{ti}\Big(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)|e_t\Big).$$

$w_i = \sum_{t=1}^{N} w_t^i$, $S_{XX',i} = \sum_{t=1}^{N} w_t^i E_{ti}(XX')$ et $S_{X,i} = \sum_{t=1}^{N} w_t^i E_{ti}(X)$.
Donc en utilisant la méthode du maximum de vraisemblance on obtient les estimateurs suivants :

$$\widehat{\mu}_i^{MV} = \frac{S_{Y,i} - B_i - S_{X,i}}{\sum_t w_t^i}$$

$$\widehat{B}_i^{MV} = (S_{YX',i} - \mu_i S_{X',i})(S_{XX',i})^{-1}.$$

Pour l'estimation de Σ_i, on désigne par $B2_i = [B_i\ \mu_i]$ et $Z = [X\ 1]$, on obtient donc

$$\widehat{\Sigma}_i^{MV} = \frac{S_{YY',i} - S_{YZ',i}B2_i' - B2_i S_{ZY',i} + B2_i S_{ZZ',i} B2_i'}{w_i}$$

1.5.3 Apprentissage de la structure

Monti et Cooper [50] proposent deux méthodologies à base de score pour l'apprentissage de structure des réseaux Bayésiens Gaussiens Conditionnels à partir de données. La première méthodologie consiste à discrétiser les données pour ensuite utiliser les algorithmes classiques applicables pour le cas discret. La deuxième méthodologie consiste à utiliser des réseaux de neurones comme estimateurs de probabilités qui seront ensuite utilisés pour calculer des scores.

1.6 Modèles linéaires Non Gaussiens

1.6.1 Définition

Un modèle LiNGAM (linéaire non gaussien acyclique) est défini par les trois propriétés suivantes :

1. Les données observées sont générées à partir d'un processus représenté graphiquement par un graphe dirigé acyclique (DAG).

2. La valeur associée à chaque variable est une fonction linéaire des valeurs associées aux variables précédentes plus un terme d'erreur e_i

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i. \tag{1.5}$$

On note par $k(i)$ l'ordre de $x(i)$.

3. Les erreurs e_i sont des variables aléatoires indépendantes de distribution non-gaussienne et de variance non nulle.

$$p(e_1, \ldots, e_n) = \prod_i p_i(e_i).$$

1.6.2 Apprentissage de la structure

Les modèles LiNGAM ne possédent pas de forme paramètriques, on s'intéresse seulement à l'apprentissage de structure.
On réécrit le modèle (1.5) sous forme matricielle comme suit :

$$x = Bx + e \tag{1.6}$$

où B est une matrice qui pourrait être permutée en une matrice triangulaire inférieure stricte si l'on connaissait un ordre topologique $k(i)$ sur les variables.
On commence d'abord par résoudre l'équation (1.6) pour x.
On obtient donc

$$x = Ae \tag{1.7}$$

où $A = (I - B)^{-1}$.
Etant donné que les composantes de e sont indépendantes et non Gaussiennes, l'équation (1.7) peut être résolue en utilisant des méthodes d'analyse en composantes indépendantes. Certaines méthodes comme Fast-ICA [32] tiennent compte de la non gaussiannité des erreurs. Ils proposent aussi différentes mesures de cette non gaussianité.
Plusieurs algorithmes ont ainsi été proposé pour estimer la structure de modèles LINGAM, ICA-LINGAM [64] basée sur l'utilisation de Fast-ICA ou DirectLINGAM [65] qui utilise des tests d'indépendance non paramètriques.

1.7 Conclusion

Dans ce chapitre, nous avons présenté les modèles graphiques probabilistes. Nous avons discuté une classe de ces MGP, notamment les différents types des RBs.
Pour chacun de ces types, nous avons étudié l'apprentissage des paramètres et de la structure.

Chapitre 2

Famille exponentielle naturelle, estimation implicite et mélange infini de lois Gaussiennes

2.1 Introduction

Les familles exponentielles naturelles (FENs) ont une longue histoire commençant avec Fisher [21, 2] et Letac [48], elles continuent à être un outil indispensable de modélisation en statistique et à fournir un cadre naturel pour certaines branches de la théorie des probabilités. Dans le cadre de l'estimation Bayésienne non informative, nous considérons l'approche implicite [26]. Cette dernière méthode a été très récemment appliquée aux RBs mais seulement aux RBs discrets [4, 5]. Les mélanges de distributions ont fourni un ensemble d'approches mathématiques pour la modélisation statistique d'une grande variété de phénomènes aléatoires. Les mélanges de distributions ont fourni un ensemble d'approches mathématiques pour la modélisation statistique d'une grande variété de phénomènes aléatoires.

Notre but est donc d'adapter cette forme paramétrique (FEN) aux RBs, d'étendre l'application de l'estimation implicite dans les cas continus et mixtes. En plus nous allons introduire la notion de mélange infini de lois Gaussiennes dans les RBs et étudier les réseaux Bayésiens associés à un modèle de mélange infini de lois gaussiennes.

Dans ce chapitre nous nous sommes intéressés à rappeler des généralités sur les familles exponentielles naturelles, à fixer les notations, à présenter la méthode implicite et à introduire les réseaux Bayésiens associés à un modèle de mélange infini de lois Gaussiennes.

2.2 Famille exponentielle naturelle

2.2.1 Généralités sur les familles exponentielles naturelles

Les familles exponentielles sont les modèles statistiques les plus utilisés en théorie classique, puisque la plupart des tests et des méthodes d'estimation des paramètres s'y appliquent. Un certain nombre de papiers ont été consacrés à la théorie des familles exponentielles naturelles (FENs) dans les dernières années. Certains de ces travaux concernent la classification des FENs. Plusieurs classifications de FENs ont été atteintes (voir [51]).
En 1982, Morris a démontré qu'il existe, à une affinitée près, six familles quadratiques (FEN dont la fonction variance V_F est égale à la restriction sur M_F d'une fonction polynomiale de degré inférieur ou égale à 2 : Normale, Poisson, Gamma, Binomiale, Binomiale Négative et Cosinus hyperbolique).
Pour une présentation exacte des familles quadratiques simples, rappelons tout d'abord quelques définitions et notions touchant les familles exponentielles naturelles qui sont devenues traditionnelles dans la statistique. Pour plus de détails, on peut se référer à [48, 2].
On désigne par $(\theta, x) \longmapsto \langle \theta, x \rangle$ le produit scalaire canonique sur $\mathbb{R}^d \times \mathbb{R}^d$ et par ν une mesure positive définie sur $\mathbb{R}^d$. L'application

$$\begin{array}{rccl} L_\nu : & \mathbb{R}^d & \longrightarrow & [0, +\infty[\\ & \theta & \longmapsto & \displaystyle\int_{\mathbb{R}^d} \exp\langle \theta, x \rangle \nu(dx) \end{array}$$

s'appelle la transformée de Laplace de la mesure ν.
On note $\Theta(\nu) = int\{\theta \in \mathbb{R}^d; L_\nu(\theta) < +\infty\}$.
On désigne par $\mathcal{M}(\mathbb{R}^d)$ l'ensemble des mesures positives ν sur $\mathbb{R}^d$ telles que ν ne soit pas concentrée sur un hyperplan affine de $\mathbb{R}^d$ et telle que $\Theta(\nu)$ soit non vide.
Pour toute mesure $\nu \in \mathcal{M}(\mathbb{R}^d)$ et pour tout $\theta \in \Theta(\nu)$, la fonction cumulante

de la mesure ν est définie par

$$k_\nu(\theta) = \log(L_\nu(\theta)).$$

Pour tout $\theta \in \Theta(\nu)$, on considère la mesure de probabilité

$$P(\theta, \nu)(dx) = e^{\langle \theta, x \rangle - k_\nu(\theta)} \nu(dx).$$

Par conséquent, la famille exponentielle naturelle (FEN) engendrée par ν est donnée par

$$F = F(\nu) = \{P(\theta, \nu)(dx);\ \theta \in \Theta(\nu)\}.$$

Il est important de signaler, grâce à l'inégalité de Hölder, que k_ν est une fonction strictement convexe et infiniment différentiable sur $\Theta(\nu)$. Sa différentielle est donc une fonction injective sur $\Theta(\nu)$, et est définie par

$$k'_\nu(\theta) = \int_{\mathbb{R}^d} xP(\theta, \nu)(dx).$$

Notons que $k'_\nu(\theta)$ est la moyenne de la probabilité $P(\theta, \nu)$ qui sera par la suite nommée la fonction des moyennes de la mesure ν associée à la FEN $F(\nu) = F$.
Si ν' et ν sont deux mesures dans $\mathcal{M}(\mathbb{R}^d)$, alors

$$F(\nu') = F(\nu) \Leftrightarrow \exists (a, b) \in \mathbb{R}^d \times \mathbb{R} \ /\ \nu'(dx) = e^{\langle a, x \rangle + b} \nu(dx).$$

On appelle domaine des moyennes de la FEN $F = F(\nu)$, l'ensemble $M_F = k'_\nu(\Theta(\nu))$.
Si ν est dans $M(\mathbb{R}^d)$, k_ν est strictement convexe et analytique réelle sur $\Theta(\nu)$. Par conséquent, k'_ν définit un difféomorphisme de $\Theta(\nu)$ sur son image M_F appelé domaine des moyennes de F. Soit $\psi_\nu : M_F \longrightarrow \Theta(\nu)$ l'application réciproque de k'_ν, nous obtenons alors une nouvelle paramétrisation de la famille F appelée paramétrisation par la moyenne : $F = \{P(\mu, F) = P(\psi_\nu(\mu), \nu); \mu \in M_F\}$ donc nous pouvons revenir à la paramétrisation canonique par $\theta = \psi_\nu(\mu)$.
Soit F une famille exponentielle naturelle, on appelle fonction variance de la famille F l'application V_F définie sur M_F par $V_F(\mu) = k''_\nu(\psi_\nu(\mu))$. Son importance est due au fait qu'elle caractérise la famille F (c'est à dire si les fonctions variance de deux FENs F_1 et F_2 coincident sur une partie ouverte

O non vide contenue dans l'intersection des domaines de moyennes de M_{F_1} et M_{F_2}, alors les deux familles F_1 et F_2 coincident).

Familles exponentielles naturelles de Morris	domaine des moyennes	fonction variance
Normale	$M_F = \mathbb{R}$	$V_F(m) = p, p > 0$
Poisson	$M_F =]0, +\infty[$	$V_F(m) = m$
Gamma	$M_F =]0, +\infty[$	$V_F(m) = \frac{m^2}{p}, p > 0$
Binomiale	$M_F =]0, N[$	$V_F(m) = m(1 - \frac{m}{N})$
Binomiale Négative	$M_F =]0, +\infty[$	$V_F(m) = m(1 + \frac{m}{p}), p > 0$
Cosinus hyperbolique	$M_F = \mathbb{R}$	$V_F(m) = p(1 + \frac{m^2}{p^2}), p > 0$

TAB. 2.1 – Familles Exponentielles naturelles de Morris

Une caractérisation des a prioris conjuguées pour le paramètre θ indexant une famille exponentielle naturelle a été fourni par Diaconis et Ylvisaker [20]. Soit $F = F(\nu)$ une FEN, et soit Π une famille de lois a priori sur $\Theta(\nu)$

$$\Pi_{t,m_0}(d\theta) = K_{t,m_0} e^{t[\langle \theta, m_0 \rangle - k_\nu(\theta)]} 1_{\Theta(\nu)}(\theta) d\theta$$

où $t > 0$, $m_0 \in M_F$ et K_{t,m_0} est la constante de normalisation. Soit X une variable aléatoire de loi $P(\theta, \nu)$, la famille Π est dite cojuguée, si la distribution a posteriori de θ sachant X appartient à Π, où la loi de (θ, X) est $\Pi_{t,m_0}(d\theta)P(\theta, \nu)(dx)$. En plus de Π, Consonni et Veronese [14] ont considéré deux autres familles de distributions a priori sur M_F. La première, $\widetilde{\Pi}$, est défini par une construction semblable quant à Π, pour un choix convenable de (t, m_0),

$$\widetilde{\Pi}_{t,m_0}(d\mu) = \widetilde{K}_{t,m_0} e^{t[\langle \psi_\nu(\mu), m_0 \rangle - k_\nu(\psi_\nu(\mu))]} 1_{M_F}(\mu) d\mu.$$

La deuxième famille de lois a priori, Π^*, est juste l'ensemble d'images de Π par k'_ν. Alors, considérons une FEN $F = F(\nu)$ réelle, Consonni et Veronesse [14] ont montré que $\widetilde{\Pi} = \Pi^*$ si et seulement si la FEN initiale $F(\nu)$ est dans la classe de Morris.

2.2.2 Estimation Bayésienne pour les familles exponentielles naturelles

Les estimateurs EAP et MAP

Soit F une FEN sur $\mathbb{R}^d$ engendrée par une mesure de probabilité ν. Soit X une variable aléatoire suivant $P(\mu, F)$ (i.e ; on note $X \sim P(\mu, F)$). On suppose que la distribution a priori de μ est $\widetilde{\Pi}_{t,m_0}$ ($\mu \sim \widetilde{\Pi}_{t,m_0}$) donc la loi conditionnelle de μ sachant $X = x$ est définie par

$$P(\mu/X = x) \propto \widetilde{K}_{t,m_0} e^{t[\langle \psi_\nu(\mu), m_0\rangle - k_\nu(\psi_\nu(\mu))]} e^{\langle \psi_\nu(\mu), x\rangle - k_\nu(\psi_\nu(\mu))} \mathbf{1}_{M_F}(\mu).$$

Ceci implique que la loi conditionnelle de μ sachant X est $\widetilde{\Pi}_{t+1, \frac{X+tm_0}{t+1}}$ et dans le cas d'un échantillon aléatoire $X_1, \ldots, X_n$, la loi conditionnelle de μ est $\widetilde{\Pi}_{t+n, \frac{\sum_{i=1}^n X_i + tm_0}{t+n}}$.
L'estimateur de moyenne a posteriori (EAP) de μ est donné par

$$\begin{aligned} \widehat{\mu}_{EAP} = E(\mu/X) &= \widetilde{K}_{t+1, \frac{X+tm_0}{t+1}} \int_{M_F} \mu e^{(t+1)[\langle \psi_\nu(\mu), \frac{X+tm_0}{t+1}\rangle - k_\nu(\psi_\nu(\mu))]} d\mu \\ &= \frac{\displaystyle\int_{M_F} \mu e^{(t+1)[\langle \psi_\nu(\mu), \frac{X+tm_0}{t+1}\rangle - k_\nu(\psi_\nu(\mu))]} d\mu}{\displaystyle\int_{M_F} e^{(t+1)[\langle \psi_\nu(\mu), \frac{X+tm_0}{t+1}\rangle - k_\nu(\psi_\nu(\mu))]} d\mu}. \end{aligned}$$

On calcule aussi l'estimateur du maximum a posteriori (MAP)

$$\begin{aligned} \widehat{\mu}_{MAP} &= \arg\max_{\mu \in M_F} \left(\widetilde{K}_{t+1, \frac{X+tm_0}{t+1}} e^{(t+1)[\langle \psi_\nu(\mu), \frac{X+tm_0}{t+1}\rangle - k_\nu(\psi_\nu(\mu))]} 1_{M_F}(\mu) \right) \\ &= \frac{X+tm_0}{t+1}. \end{aligned}$$

Dans le cas d'un échantillon aléatoire $X_1, \ldots, X_n$ de loi $P(\mu, F)$, l'estimateur MAP de μ est

$$\widehat{\mu}_{MAP} = \frac{\displaystyle\sum_{i=1}^n X_i + tm_0}{t+n}.$$

Si $X_1, \ldots, X_n$ est un échantillon aléatoire de loi $P(\mu, F)$, d'après la loi des grands nombres $\overline{X}_n = \frac{1}{n}\displaystyle\sum_{i=1}^n X_i \overset{P-S}{\longrightarrow} \mu \in M_F$. Alors l'estimateur du maximum de vraisemblance $\widehat{\mu}_{MV} = \overline{X}_n$ existe si n est assez grand.

Exemples

On conclut cette partie en donnant des exemples de FENs : Poisson, Binomiale et Binomiale Négative. Pour chaque famille nous commençons d'abord par spécifier $k_\nu(\theta), \psi_\nu(\mu)$ et M_F.

• Famille de lois de Poisson : Pour cette famille, $X \sim P(\mu, \nu) = \mathcal{P}(\mu)$ (i.e $\nu = \sum_{k\in\mathbb{N}} \frac{\delta_k}{k!}$ *et* $P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$), la fonction cumulante est $k_\nu(\theta) = e^\theta, \Theta(\nu) = \mathbb{R}$, le domaine des moyennes de F est $M_F =]0, +\infty[$ et $\theta = \psi_\nu(\mu) = \log(\mu)$. Nous vérifions que la constante de normalisation est

$$\widetilde{K}_{t,m_0} = \frac{1}{\displaystyle\int_{M_F} e^{t(\psi_\nu(\mu)m_0-\mu)}d\mu} = \frac{t^{tm_0+1}}{\Gamma(tm_0+1)}$$

où Γ est la fonction Gamma.

La famille $\widetilde{\Pi}$ des distributions a priori sur M_F est

$$\widetilde{\Pi}_{t,m_0}(d\mu) = \frac{t^{tm_0+1}}{\Gamma(tm_0+1)} e^{t(m_0\log(\mu)-\mu)} 1_{]0,+\infty[}(\mu)d\mu, \quad t > 0 \text{ et } m_0 \in M_F.$$

L'estimateur EAP de μ est

$$\begin{aligned}
\widehat{\mu}_{EAP} = E(\mu/X_1, \ldots, X_n) &= \widetilde{K}_{t+1,\frac{X+tm_0}{t+1}} \int_0^{+\infty} \mu e^{(t+1)[\log(\mu)\frac{X+tm_0}{t+1}]-(t+1)\mu} d\mu \\
&= \widetilde{K}_{t+1,\frac{X+tm_0}{t+1}} \int_0^{+\infty} \mu^{X+tm_0+1} e^{-(t+1)\mu} d\mu \\
&= \frac{(t+1)^{X+tm_0+1}}{\Gamma(X+tm_0+1)} \frac{\Gamma(X+tm_0+2)}{(t+1)^{X+tm_0+2}} \\
&= \frac{X+tm_0+1}{t+1}.
\end{aligned}$$

Dans le cas d'un échantillon $X_1, \ldots, X_n$ de loi de Poisson $\mathcal{P}(\mu)$, l'estimateur EAP de μ est

$$\frac{\displaystyle\sum_{i=1}^n X_i + tm_0 + 1}{t+n}.$$

• Famille de lois Binomiale Négative : Pour cette famille $X \sim P(\mu, \nu) = NB(\mu, 1)$, la fonction cumulante est $k_\nu(\theta) = -\log(1-e^\theta)$, $\Theta(\nu) =]-\infty, 0[$,

le domaine des moyennes de F est $M_F =]0, +\infty[$ et $\psi_\nu(\mu) = \log(\frac{\mu}{\mu+1})$.
On prouve que la constante de normalisation est

$$\widetilde{K}_{t,m_0} = \frac{1}{\displaystyle\int_{M_F} e^{t(\log(\frac{\mu}{\mu+1})m_0 + \log(1-\frac{\mu}{\mu+1}))} d\mu} = \frac{\Gamma(tm_0+t)}{\Gamma(tm_0+1)\Gamma(t-1)},\ t > 1.$$

Ainsi, la famille de lois a priori sur M_F est

$$\widetilde{\Pi}_{t,m_0}(d\mu) = \frac{\Gamma(tm_0+t)}{\Gamma(tm_0+1)\Gamma(t-1)} e^{t\log(\frac{\mu}{\mu+1})m_0 + t\log(1-\frac{\mu}{\mu+1})} 1_{]0,+\infty[}(\mu) d\mu.$$

L'estimateur EAP de μ est

$$\begin{aligned}
\widehat{\mu}_{EAP} = E(\mu/X) &= \widetilde{K}_{t+1,\frac{X+tm_0}{t+1}} \int_0^{+\infty} \mu e^{\log(\frac{\mu}{\mu+1})(tm_0+X) + (t+1)\log(1-\frac{\mu}{\mu+1})} d\mu \\
&= \widetilde{K}_{t+1,\frac{X+tm_0}{t+1}} \int_0^{+\infty} \mu (\frac{\mu}{\mu+1})^{tm_0+X} (1-\frac{\mu}{\mu+1})^{(t+1)} d\mu.
\end{aligned}$$

En utilisant le changement de variable ($z = \frac{\mu}{\mu+1}$), on obtient

$$\begin{aligned}
\widehat{\mu}_{EAP} &= \widetilde{K}_{t+1,\frac{X+tm_0}{t+1}} \int_0^1 z^{tm_0+X+1}(1-z)^{t-2} dz \\
&= \frac{\Gamma(X+tm_0+t+1)}{\Gamma(X+tm_0+1)\Gamma(t)} \frac{\Gamma(tm_0+X+2)\Gamma(t-1)}{\Gamma(tm_0+X+t+1)} \\
&= \frac{tm_0+X+1}{t-1}.
\end{aligned}$$

Dans le cas où $(X_1, \ldots, X_n)$ est un n-échantillon aléatoire de loi $NB(\mu, 1)$, l'estimateur EAP de μ est donné par

$$\widehat{\mu}_{EAP} = E(\mu/X_1, \ldots, X_n) = \frac{\displaystyle\sum_{i=1}^{n} X_i + tm_0 + 1}{t+n-2}.$$

• Famille de lois Binomiale : $X \sim P(\mu, \nu) = B(N, p = \frac{\mu}{N})$ (i.e $P(X = k) = C_N^k (\frac{\mu}{N})^k (1-\frac{\mu}{N})^{N-k}$ avec $k \in \{0, 1, \ldots, N\}$, la fonction cumulante est $k_\nu(\theta) = N\log(1+e^\theta), \Theta(\nu) =]-\infty, 0[$, le domaine des moyennes de F est $M_F =]0, N[$ et $\psi_\nu(\mu) = \log(\frac{\mu}{N-\mu})$, après un calcul on obtient

$$\widetilde{K}_{t,m_0} = \frac{\Gamma(tN+2)}{N\Gamma(tm_0+1)\Gamma(tN-tm_0+1)}.$$

On remarque que $p = \frac{\mu}{N} \sim Beta(tm_0+1, tN-tm_0+1)$ et la loi a posteriori de p sachant X est $Beta(X+tm_0+1, (t+1)N-X-tm_0+1)$ donc

$$\widehat{p}_{EAP} = \frac{\widehat{\mu}}{N} = E(\frac{\mu}{N}/X) = \frac{X+tm_0+1}{(t+1)N+2}$$

ce qui correspond au cas classique.
Le théorème suivant étudie le comportement asymptotique de l'estimateur EAP $\widehat{\mu}_n$ de μ dans le cas réel.

Théorème 2.1. *1)* $\widehat{\mu}_n \overset{n\to+\infty}{\longrightarrow} \mu$ *presque sûrement.*
2) $\sqrt{n}(\widehat{\mu}_n - \mu) \overset{n\to+\infty}{\longrightarrow} N(0, V_F(\mu))$ *en loi.*

2.3 Estimation implicite

Dans le contexte de la théorie Bayésienne [61, 6], le paramètre inconnu θ dans le modèle statistique $\{P_\theta(dx)\,;\,\theta \in \Theta\}$ est supposé une variable aléatoire de loi connue. Cette loi a priori, que l'on pourrait traduire par la connaissance préliminaire que l'on a du problème, est modifiée à la lumière des données pour déterminer une loi a posteriori (la loi conditionnelle de θ sachant les données). Cette loi qui résume à la fois ce que l'on peut dire de θ à partir des hypothèses a priori, et ce qu'apportent les données. Cette méthode permet de tirer quelques conclusions, mais elle dépend des hypothèses qui pourraient être prises dans diverses conditions. Ainsi, l'idée de base de la théorie Bayésienne est de considérer le paramètre θ comme une variable aléatoire et de déterminer la loi conditionnelle de ce paramètre sachant les données. De ce point de vue, la loi implicite [26] peut jouer le rôle d'une loi a posteriori dans la méthode Bayésienne. Parce qu'elle est obtenue dans un contexte naturel sans spécifier de nouvelles hypothèses ou des lois a priori.
Pour expliquer le principe de la loi implicite, on considère une famille de lois de probabilités $\{P_\theta(dx) = P(x/\theta)\nu(dx), \theta \in \Theta\}$ paramétrisée par un paramètre inconnu θ dans un ensemble Θ, où x est une réalisation d'une variable aléatoire X.
La loi implicite $P(\theta/x)$ est calculée en multipliant la fonction de vraisemblance $P(x/\theta)$ par la mesure de comptage σ si Θ est un ensemble discret et par la mesure de Lebesgue σ si Θ est un ensemble ouvert (σ dépend seulement de la structure topologique de Θ) et en divisant par la constante de normalisation $c(x) = \displaystyle\int_\Theta P(x/\theta)\sigma(d\theta)$. Par conséquent la loi implicite est donnée par la formule suivante $P(\theta/x) = (c(x))^{-1}P(x/\theta)\sigma(\theta)$. L'estimateur implicite $\widehat{\theta}$ de θ est simplement la moyenne de la loi implicite.

Modèle Multinomial :

Soit $X = (N_1, \ldots, N_r)$ un vecteur aléatoire suivant la loi multinomiale $Mult(N, \theta_1, \ldots, \theta_r)$ de paramètres $N = \sum_{i=1}^{r} N_i$ et $(\theta_1, \ldots, \theta_r) \in]0,1[^r$ tels que $\sum_{i=1}^{r} \theta_i = 1$;

$$P(X/\theta) = P(N_1 = n_1, \ldots, N_r = n_r) = \frac{N!}{n_1! \ldots n_r!} \theta_1^{n_1} \ldots \theta_r^{n_r}.$$

La loi implicite de θ sachant $X = (N_1, \ldots, N_r)$ est une distribution de Dirichlet de paramètres $N_1 + 1, \ldots, N_r + 1$, notée $Dir(N_1 + 1, \ldots, N_r + 1)$.
La distribution de probabilité de l'observation suivante à l'état k est alors donnée par : $\forall\ k \in \{1, \ldots, r\}$

$$\begin{aligned} \widehat{\theta}_k &= \int \theta_k Dir(N_1 + 1, \ldots, N_r + 1)(d\theta) \\ &= \frac{N_k + 1}{N + r}. \end{aligned}$$

$\widehat{\theta}_k$ est l'estimateur implicite de θ_k. On remarque que l'estimateur implicite correspond, dans le cas où N est connu, à l'estimateur Bayésien avec toutes les α_i égales à 1 (a priori uniforme).

Modèle Gaussien :

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $N(0, \lambda)$. La fonction vraisemblance de λ pour n-observations $\underline{x} = (x_1, \ldots, x_n)$ est définie par

$$L(x_1, \ldots, x_n, \lambda) = (2\pi\lambda)^{-\frac{n}{2}} \exp\Big(- \frac{1}{2\lambda} \sum_{i=1}^{n} x_i^2 \Big).$$

La loi implicite $Q_{\underline{x}}$ de λ est égale à

$$Q_{\underline{x}}(d\lambda) = \frac{\lambda^{-\frac{n}{2}} \exp^{-\frac{\sum_{i=1}^{n} x_i^2}{2\lambda}}}{\Gamma(\frac{n}{2} - 1)} \Big(\frac{1}{2} \sum_{i=1}^{n} x_i^2 \Big)^{\frac{n}{2} - 1}.$$

L'estimateur implicite de λ est défini par : $\forall n \geq 5$

$$T_n = E(\lambda / X_1, \ldots, X_n) = \frac{1}{n-4} \sum_{i=1}^{n} X_i^2.$$

2.4 Mélange infini

Vue l'utilité des mélanges de lois de probabilités comme méthode de modélisation, ces modèles de mélange ont continué à avoir un succès et un intérêt croissant du point de vue pratique et théorique. Les domaines dans lesquels les modèles de mélange ont été appliqués avec succès comprennent l'astronomie, la biologie, la génétique, la médecine, la psychiatrie, l'économie, l'ingénierie et le marketing. Dans les dernières années, les modèles de mélanges finis ont reçu une prise en compte croissante et ont été beaucoup étudiés à la fois théorique et pratique par plusieurs auteurs Mclachlan et Peel [49]. Les mélanges finis de distributions ont fourni une approche mathématique fondée sur la modélisation statistique d'une grande variété de phénomènes aléatoires. De nos jours, ces modèles sont nécessaires dans de nombreuses applications qui se posent dans les différents domaines de connaissance. Les modèles de mélanges finis ont été largement utilisés dans la littérature. La densité du mélange fini de lois Gaussiennes de X est définie par

$$f_X(x) = \sum_{z=1}^{K} \pi_z N(\mu(z), \Sigma(z))(x)$$

où $N(\mu(z), \Sigma(z))$ est une distribution normale de moyenne $\mu(z)$ et de matrice de variance covariance $\Sigma(z)$, et les proportions de mélange $0 < \pi_z < 1$ satisfait $\sum_{z=1}^{K} \pi_z = 1$.
On suppose que la loi conditionnelle de X sachant $Z = z$ est une loi normale centrée de variance $\sigma^2(z)$ (i.e ; $X/Z = z \sim N(0, \sigma^2(z))$ telle que Z est une variable aléatoire de densité $f(z)$. La densité du mélange infini de X est donnée par

$$\begin{aligned} f_X(x) &= \int_0^{\infty} f_{(X/Z=z)}(x) f(z) dz \\ &= \int_0^{\infty} N(\mu(z), \sigma^2(z))(x) f(z) dz \end{aligned}$$

où

$$\begin{aligned} \mu(z) &= E(X/Z = z) \\ \sigma^2(z) &= var(X/Z = z). \end{aligned}$$

On note, si la distribution de X sachant $Z = z$ est Normale $N(\mu, z^{-1}B)$ et la distribution de Z est Gamma $\gamma(\frac{\nu}{2}, \frac{\nu}{2})$ alors X suit la loi de Student

de moyenne μ, de degré de liberté ν et de variance $\Sigma = \frac{\nu}{\nu-2}B$ $(i.e; X \sim t(\mu, B^{-1}, \nu))$.
En effet, il suffit de prouver que la densité de probabilité de X peut être réécrite comme

$$t(x/\mu, B^{-1}, \nu) = \int_0^\infty N(x/\mu, z^{-1}B)\gamma(z/\frac{\nu}{2}, \frac{\nu}{2})dz.$$

Nous pouvons exprimer la distribution de student hiérarchiquement comme

$$X/(\mu, B^{-1}, \nu) \sim N(\mu, z^{-1}B)$$

$$Z/\nu \sim \gamma(\frac{\nu}{2}, \frac{\nu}{2}).$$

On prouve le résultat suivant

Proposition 2.1. *Soient X et Z deux variables aléatoires telles que la loi conditionnelle de X sachant $Z = z$ est $N(0, \sigma^2(z))$. Le mélange infini de X est Normale si et seulement si X et Z sont indépendantes.*

Preuve 2.1. " $\Longrightarrow$ " *On suppose que le mélange infini de X est normale de moyenne $\mu = 0$ et de variance σ^2.*
$\sigma^2 = var(X) = \int_0^\infty \sigma^2(z)f(z)dz.$
Etant donné que $X \sim N(0, \sigma^2)$ alors $E(e^{i\theta X}) = e^{-\frac{\sigma^2\theta^2}{2}}, \forall\theta \in \mathbb{R}$.
Comme

$$\varphi_X(\theta) = E(e^{i\theta X}) = \int_{\mathbb{R}} e^{i\theta x} \int_0^\infty N(0, \sigma^2(z))(x)f(z)dzdx$$

alors

$$\varphi_X(\theta) = E(e^{i\theta X}) = \int_0^\infty e^{-\frac{\sigma^2(z)\theta^2}{2}} f(z)dz = e^{-\frac{\sigma^2\theta^2}{2}}.$$

En prenant la différentielle par rapport à θ, on aura

$$\varphi'_X(\theta) = -\int_0^\infty [\theta\sigma^2(z)]e^{-\frac{\sigma^2(z)\theta^2}{2}} f(z)dz.$$

La dérivée de φ_X à l'ordre 4 est donnée par

$$\varphi_X^{(4)}(\theta) = \int_0^\infty f(z)\{3\sigma^4(z) + 6\sigma^2(z)(\theta\sigma^2(z))^2 + (\theta\sigma^2(z))^4\}e^{-\frac{\sigma^2(z)}{2}\theta^2}dz.$$

Pour $\theta = 0$

$$\varphi_X^{(4)}(0) = 3\int_0^\infty f(z)\sigma^4(z)dz = 3\sigma^4,$$

alors

$$\sigma^4 = (\sigma^2)^2 = (\int_0^\infty f(z)\sigma^2(z)dz)^2 = \int_0^\infty f(z)\sigma^4(z)dz. \qquad (2.1)$$

D'autre part, en utilisant l'inégalité de Hölder, nous avons

$$\Big(\int_0^\infty 1\sigma^2(z)f(z)dz\Big)^2 \leq \Big(\int_0^\infty \sigma^4(z)f(z)dz\Big)\Big(\int_0^\infty 1f(z)dz\Big)^2 = \int_0^\infty \sigma^4(z)f(z)dz. \qquad (2.2)$$

Dans notre cas, l'équation (2.2) devient une égalité qui signifie qu'il existe $\alpha > 0$ *tel que* $\sigma^2(z) = \alpha \times 1$ *est constante. Nécessairement* $\sigma^2(z) = \sigma^2$.
Par conséquent, la distribution conditionnelle de $X/Z = z$ *est indépendante de* Z. *Donc* X *et* Z *sont indépendantes.*
" $\Longleftarrow$ " *évident*

Dans la suite, on suppose que $Z \sim \gamma(\nu, \lambda)$ (i.e $f_Z(z) = \frac{\lambda^\nu}{\Gamma(\nu)} e^{-\lambda z} z^{\nu-1} \mathbf{1}_{]0,+\infty[}(z)$), telle que $X/Z = z \sim N(\mu(z), \sigma^2(z))$ de telle sorte que $\mu(z) = \frac{a}{z} + b, \sigma^2(z) = \frac{\sigma^2}{z}$ et $a \geq 0, b \in \mathbb{R}, \sigma^2 > 0$.
La loi de X est appelée loi Gamma Normale infinie et notée $X \sim N_\infty(a, b, \sigma^2, \nu, \lambda)$.

Proposition 2.2. *Si* $X \sim N_\infty(a, b, \sigma^2, \nu, \lambda)$ *alors la fonction densité de* X *est donnée par*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\lambda^\nu}{\Gamma(\nu)} e^{\frac{xa-ab}{\sigma^2}} K(\nu + \frac{1}{2}, \lambda + \frac{(x-b)^2}{2\sigma^2}, \frac{a^2}{2\sigma^2})$$

où K *désigne la fonction de Bessel modifiée définie par*

$$K(\nu, a, b) = \int_0^\infty e^{-(ax + \frac{b}{x})} x^{\nu-1} dx$$

La proposition suivante permet de déterminer la fonction caractéristique de la loi $N_\infty(a, b, \sigma^2, \nu, \lambda)$.

Proposition 2.3. *Si* $X \sim N_\infty(a, b, \sigma^2, \nu, \lambda)$ *alors la fonction caractéristique de* X *est donnée par*

$$\varphi_X(t) = E(e^{itX}) = e^{ibt} \frac{\lambda^\nu}{\Gamma(\nu)} K(\nu, \lambda, \frac{\sigma^2 t^2}{2} + ita).$$

Preuve 2.2. *La fonction caractéristique de* X *est définie par*

$$\begin{aligned} \varphi_X(t) &= \int_0^\infty \varphi_{(X/Z=z)}(t) f_Z(z) dz \\ &= \frac{\lambda^\nu}{\Gamma(\nu)} \int_0^\infty e^{-\frac{\sigma^2}{2z} t^2 + i(\frac{a}{z} + b)t} e^{-\lambda z} z^{\nu-1} dz \\ &= e^{ibt} \frac{\lambda^\nu}{\Gamma(\nu)} K(\nu, \lambda, \frac{\sigma^2 t^2}{2} + ita). \end{aligned}$$

La proposition suivante montre que le mélange infini N_∞ est stable par une transformation affine.

Proposition 2.4. *Si* $X \sim N_\infty(a, b, \sigma^2, \nu, \lambda)$ *alors* $\alpha X + \beta \sim N_\infty(a\alpha, b\alpha + \beta, \alpha^2\sigma^2, \nu, \lambda)$

Théorème 2.2. *Le modèle de mélange infini* $M_\infty = \{N_\infty(a, b, \sigma^2, \nu, \lambda); a \geq 0, b \in \mathbb{R}, \sigma^2 > 0, \nu > 2, \lambda = \lambda_0$ *connu*$\}$ *est identifiable.*

La preuve de ce théorème nécessite le lemme suivant qui est dû à Bruni et Koch [8]. Dans le cas réel, soit D un compact de $\mathbb{R}$. Soit $\theta(z) = (\theta_1, \theta_2) = (\mu(z), \sigma^2(z))$ désigne la fonction à deux variables définie sur D, dont les composantes représentent, respectivement, la fonction de valeur moyenne μ et de fonction de variance σ^2 de la famille de densité Gaussienne. Nous introduisons maintenant des ensembles de fonctions θ. Le premier ensemble est défini comme suit :

$$\Lambda_1 = \Big\{\theta \in C^1(D) : |\theta_1(z)| \leq K_1 < \infty, 0 < s_1 \leq \theta_2(z) \leq s_2 < \infty,$$

$$|\dot{\theta}_1(z)| + |\dot{\theta}_2(z)| \leq K_2 < \infty, \forall z \in D\Big\}.$$

Un deuxième ensemble est

$$\Lambda_2 = \Big\{\theta \in \Lambda_1 : \theta(z) = \theta(z') \Longrightarrow z = z' \text{ et } \theta_1, \theta_2 \text{ sont monotones}$$

$$\text{et } |\dot{\theta}_1(z)| + |\dot{\theta}_2(z)| > 0, \forall z \in D\Big\}.$$

On considère $f = T(\theta, \gamma), f(x) = \displaystyle\int_D N(x; \theta(z))\gamma(dz)$

Lemme 2.1. *Si l'on entend par* $M(D)$ *l'ensemble des mesures signées de support* D*, on notera* $\mathcal{P}(D)$ *le sous ensemble des probabilités de* $M(D)$*. Etant donné* $\theta, \theta' \in \Lambda_2, \gamma$ *et* $\gamma' \in \mathcal{P}(D)$*, si* $T(\theta, \gamma) = T(\theta', \gamma')$ *alors il existe une fonction* h *strictement monotone de classe* C^1 *de* D *à valeurs dans* D*, définie uniquement sur le support de* γ*, telle que*

$$\gamma'(dh(z)) = \gamma(dz) \text{ et } \theta'(h(z)) = \theta(z).$$

Preuve du Théorème 2.2 Soit $D \subset]0, +\infty[$
$z \longmapsto \mu(z) = \frac{a}{z} + b$ et $z \longmapsto \frac{\sigma^2}{z}$
Comme $(\mu, \sigma^2) \in \Lambda_1$, alors

$$\begin{aligned}\mu(z) &= \mu(z')\\ \sigma^2(z) &= \sigma^2(z')\end{aligned}$$

Donc $z = z'$

et $|\dot{\mu}(z)| + |\dot{\sigma}^2(z)| > 0 \ \ \forall z \in D$ alors $(\mu, \sigma^2) \in \Lambda_2$

On suppose que $N_\infty(a, b, \sigma^2, \lambda_0, \nu) = N_\infty(a', b', \sigma'^2, \lambda_0, \nu')$. Selon le lemme précédent $\exists : h : D \longrightarrow D$ une fonction strictement monotone de classe C^1 telle que $\forall z \in D$

$$\begin{cases} \gamma(\nu, \lambda)(z) & = |(h^{-1})'(z)|\gamma(\nu', \lambda')(h^{-1}(z)) \\ \frac{a}{z} + b & = \frac{a'}{h(z)} + b' \\ \frac{\sigma^2}{z} & = \frac{\sigma'^2}{h(z)}. \end{cases}$$

Par conséquent $h(z) = \frac{\sigma'^2 z}{\sigma^2}$. Ceci implique que

$$\begin{cases} \frac{a}{\sigma^2} & = \frac{a'}{\sigma'^2} \\ b & = \text{b'} \\ \nu & = \nu' \\ \frac{\lambda_0}{\sigma^2} & = \frac{\lambda_0}{\sigma'^2} \end{cases}$$

On obtient donc $\sigma^2 = \sigma'^2$ et $a = a'$.

La proposition suivante détermine la moyenne et la variance de la loi Gamma Normale infinie.

Proposition 2.5. *Si $X \sim N_\infty(a, b, \sigma^2, \nu, \lambda)$ alors*

$$E(X) = \frac{a\lambda_0}{\nu - 1} + b \text{ et } var(X) = \frac{\lambda_0}{\nu - 1}\sigma^2 + \frac{a^2\lambda_0^2}{(\nu - 1)^2(\nu - 2)}.$$

Proposition 2.6. *Si $X \sim N_\infty(a, b, \sigma^2, \nu, \lambda_0)$ alors $\epsilon = X - b - \frac{a\lambda_0}{\nu-1} \sim N_\infty(a, -\frac{a\lambda_0}{\nu-1}, \sigma^2, \nu, \lambda_0)$ telle que $E(\epsilon) = 0$ et $var(\epsilon) = \frac{\sigma^2\lambda_0}{\nu-1} + \frac{a^2\lambda_0^2}{(\nu-1)^2(\nu-2)}$.*

Théorème 2.3. *Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire tel que $X/Z = z \sim N(\frac{a}{z} + b, \frac{\Sigma}{z})$ et $Z \sim \gamma(\nu, \lambda_0)$ avec $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$ et Σ est une matrice symétrique définie positive. Alors*

1) $X \sim N_\infty(a, b, \Sigma, \nu, \lambda_0)$

2) $E(X) = \frac{a\lambda_0}{\nu-1} + b$

3) $var(X) = \Sigma_\infty = \frac{\lambda_0}{\nu-1}\Sigma + \frac{a \otimes a\lambda_0^2}{(\nu-1)^2(\nu-2)}$ (i.e ; $a \otimes a = a^t a$)

4) $CX + D \sim N_\infty(Ca, Cb + D, C\Sigma\ {}^tC, \nu, \lambda_0)$

où C est une matrice de taille $m \times n$ et D est un vecteur colonne de $\mathbb{R}^n$.

Chapitre 3

Réseau Bayésien exponentiel discret

3.1 Introduction

Dans la littérature on trouve beaucoup de travaux sur les réseaux Bayésiens discrets [27, 45, 29] où la distribution conditionnelle de chaque variable sachant ses parents est une distribution multinomiale.
Dans ce chapitre, on s'intéresse à élargir la distribution des variables à une famille exponentielle naturelle (FEN) qui représente une classe très importante de distributions dans la théorie de probabilités et statistique.
On introduit une famille de distributions a priori qui généralise l'a priori Dirichlet utilisée dans le réseau Bayésien discret et on détermine la distribution a postériori globale. On développe les estimateurs Bayésiens des paramètres et une nouvelle fonction score qui étend le score Dirichlet Bayésien pour l'apprentissage de structure.
Nous avons déterminé de façon empirique dans quels contextes ces modèles (les réseaux Bayésiens exponentiels discrets) peuvent être une bonne alternative par rapport aux réseaux Bayésiens usuels pour l'estimation de densité. Nos résultats sont illustrés par une étude de simulation.

3.2 Réseau Bayésien exponentiel discret

Soit F une FEN engendrée par une mesure de probabilité $\nu \in \mathcal{M}(\mathbb{R}^d)$.
Dans cette section, et dans ce qui suit, on note par $\Theta(\nu) = \Theta, k_\nu = k, \psi_\nu = \psi$.
Soit $X_1, \ldots, X_n$ n variables aléatoires. On suppose que $X_i/(pa_i = j) \sim P(\mu_{ij}, F)$ où $pa_i = pa(X_i)$ et $\mu_{ij} \in M_F$.
On note $\mu_i = (\mu_{ij})_{1 \leq j \leq q_i}$, $\mu = (\mu_i)_{1 \leq i \leq n}$ et $\mu_{ij} \sim \widetilde{\Pi}_{t_{ij}, m_{ij}}$.
On suppose que $(\mu_{ij})_{1 \leq j \leq q_i, 1 \leq i \leq n}$ sont indépendantes.

3.2.1 Définition

Le réseau Bayésien exponentiel discret (RBed) est défini comme un RB où les lois de probabilités conditionnelles appartiennent à des FENs. Pour ce travail, nous restreignons aux FENs discrètes. Nous appliquons les résultats précédents aux échantillons de réseau Bayésien exponentiel discret.
Nous supposons que nous avons un ensemble de données $d = \{x^{(1)}, \ldots, x^{(M)}\}$ de taille M où $x^{(h)} = \{x_1^{(h)}, \ldots, x_n^{(h)}\}$ est le h^{eme} échantillon et $x_i^{(h)}$ est la valeur de la variable X_i pour cet échantillon.
Donc, chaque distribution de probabilité conditionnelle peut être exprimée d'une façon "exponentielle"

$$P(x_i^{(h)}/pa_i^{(h)} = j, \mu, G) = e^{\langle \psi(\mu_{ij}), x_i^{(h)} \rangle - k(\psi(\mu_{ij}))} \nu\{x_i^{(h)}\}.$$

3.2.2 Généralisation de l'a priori de Dirichlet

La distribution a priori utilisé pour les RBs discrets est une Dirichlet. Nous proposons dans cette section une généralisation de cette a priori pour les RBs exponentiels discrets.
Pour la famille exponentielle naturelle discrète, nous utilisons la fonction d'a priori défini par

$$\widetilde{\Pi}_{t_{ij}, m_{ij}}(\mu_{ij}) = \widetilde{K}_{t_{ij}, m_{ij}} e^{t_{ij} < \psi(\mu_{ij}), m_{ij} > - t_{ij} k(\psi(\mu_{ij}))}$$

Prouvons que cette a priori est une distribution de Dirichlet lorsque la distribution initiale est une multinomiale.
Soit F une FEN multinomiale sur $\mathbb{R}^d$ engendrée par ν de la fonction cumu-

lante définie par

$$k_\nu(\theta) = N \log(1 + \sum_{i=1}^{d} e^{\theta_i})$$

et

$$k'_\nu(\theta) = \Big(\frac{Ne^{\theta_1}}{1 + \sum_{i=1}^{d} e^{\theta_i}}, \ldots, \frac{Ne^{\theta_d}}{1 + \sum_{i=1}^{d} e^{\theta_i}}\Big).$$

La fonction réciproque de $k'_\nu(\mu)$ est

$$\psi_\nu(\mu) = \Big(\log\Big(\frac{\mu_1}{N - \sum_{i=1}^{d} \mu_i}\Big), \ldots, \log\Big(\frac{\mu_d}{N - \sum_{i=1}^{d} \mu_i}\Big)\Big);$$

telle que

$$\mu \in \{(\mu_1, \ldots \mu_d) \in (0,1)^d; \sum_{i=1}^{d} \mu_i < N\}.$$

Si $\mu \sim \widetilde{\Pi}_{t,m_0}$ alors $\frac{\mu}{N}$ suit une loi de Dirichlet

$$\widetilde{\Pi}_{t,m_0} = Dir(tm_1 + 1, \ldots, tm_d + 1, tN - t\sum_{i=1}^{d} m_i + 1).$$

En fait, la constante de normalisation est

$$\widetilde{K}_{t,m_0} = \frac{\Gamma(tN + d + 1)}{\Gamma(tm_1 + 1) \ldots \Gamma(tm_d + 1)\Gamma(tN - t\sum_{i=1}^{d} m_i + 1)}.$$

Ce qui prouve que la famille des a prioris $\widetilde{\Pi}$ est une généralisation des a prioris Dirichlet pour la famille multinomiale.

3.2.3 Apprentissage de la structure

Dans cette section nous nous intéressons à étendre le calcul de $P(d/G)$ pour le RBed et nous proposons une fonction de score qui généralise le score Dirichlet Bayésien (BD). Nous notons cette nouvelle fonction de score par gBD (Dirichlet Bayésien généralisé).
La proposition suivante donne la probabilité de la base de données d sachant μ et la structure G.

Proposition 3.1. *On suppose qu'on a les conditions de la définition 3.2.1, donc*

$$P(d/\mu, G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} e^{\langle \psi(\mu_{ij}),\ \sum\limits_{h\in M_{ij}} x_i^{(h)}\rangle - N_{ij}k(\psi(\mu_{ij}))} \prod_{h\in M_{ij}} \nu\{x_i^{(h)}\}.$$

où $M_{ij} = \{h \in \{1,\ldots,M\}/pa_i^{(h)} = j\}$ *et* $N_{ij} = |M_{ij}|$.

Preuve 3.1.

$$\begin{aligned} P(d/\mu, G) &= \prod_{i=1}^{n}\prod_{h\in M_{ij}} P(x_i^{(h)}/pa_i^{(h)}, \mu_i, G) \\ &= \prod_{i=1}^{n}\prod_{h\in M_{ij}}\prod_{j=1}^{q_i} P(x_i^{(h)}/pa_i^{(h)} = j_i(d), \mu_{ij}, G) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{q_i} e^{\langle \psi(\mu_{ij}),\ \sum\limits_{h\in M_{ij}} x_i^{(h)}\rangle - N_{ij}k(\psi(\mu_{ij}))} (\prod_{h\in M_{ij}} \nu\{x_i^{(h)}\}) \end{aligned}$$

où $j = j_i(d)$ *désigne l'état des parents* pa_i *de* X_i.

Dans le théorème suivant nous développons une nouvelle fonction score gBD(d,G) du réseau Bayésien exponentiel discret qui est proportionnel à $P(d/G)$.

Théorème 3.1. *Supposons qu'on ait, les conditions de la section 3.2, alors*

$$P(d/G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{\widetilde{K}_{t_{ij},m_{ij}}}{\widetilde{K}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}}{N_{ij}+t_{ij}}}} \prod_{h\in M_{ij}} \nu\{x_i^{(h)}\} \qquad (3.1)$$

et

$$gBD(d,G) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{\widetilde{K}_{t_{ij},m_{ij}}}{\widetilde{K}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}}{N_{ij}+t_{ij}}}}.$$

Preuve 3.2.

$$
\begin{aligned}
P(d/G) &= \prod_{i=1}^{n} \int_{\mu_i} \prod_{h \in M_{ij}} P(x_i^{(h)}/pa_i^{(h)}, \mu_i, G) \widetilde{\Pi}(\mu_i) d\mu_i \\
&= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int_{\mu_{ij} \in M_F} e^{\langle \psi(\mu_{ij}), \sum\limits_{h \in M_{ij}} x_i^{(h)} \rangle - N_{ij} k(\psi(\mu_{ij}))} \\
&\quad \widetilde{K}_{t_{ij}, m_{ij}} e^{t_{ij} \langle \psi(\mu_{ij}), m_{ij} \rangle - t_{ij} k(\psi(\mu_{ij}))} d\mu_{ij} \prod_{h \in M_{ij}} \nu\{x_i^{(h)}\} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int_{\mu_{ij} \in M_F} e^{\langle \psi(\mu_{ij}), \sum\limits_{h \in M_{ij}} x_i^{(h)} + t_{ij} m_{ij} \rangle - (N_{ij} + t_{ij}) k(\psi(\mu_{ij}))} \\
&\quad \widetilde{K}_{t_{ij}, m_{ij}} d\mu_{ij} \prod_{h \in M_{ij}} \nu\{x_i^{(h)}\} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\widetilde{K}_{t_{ij}, m_{ij}}}{\widetilde{K}_{N_{ij}+t_{ij}, \frac{\sum_{h \in M_{ij}} x_i^{(h)} + t_{ij} m_{ij}}{N_{ij} + t_{ij}}}} \prod_{h \in M_{ij}} \nu\{x_i^{(h)}\}.
\end{aligned}
$$

Décomposabilité

La formule (3.1) se décompose bien en un produit sur chaque variable d'une fonction locale ne dépendant que du noeud i et ses parents. Ce score est donc décomposable.

Score équivalence

Nous avons essayé de reprendre la démonstration de cette propriété effectuée dans le cas multinomial (Section 1.3.3), sans réussir à prouver cette propriété pour cette nouvelle fonction de score. Une étude expérimentale non reportée ici et utilisant l'implémentation faite en section 3.4 nous a montré que deux structures équivalentes au sens de Markov avaient le même score seulement lorsque $t_{ij} = m_{ij} = 0$.

3.2.4 Apprentissage des paramètres

Le théorème suivant détermine la distribution a posteriori globale de μ sachant d.

Théorème 3.2. *Supposons qu'on ait, les conditions de la section 3.2, alors les marginales de $(\mu_i)_{1 \leq i \leq n}$ sont mutuellement indépendants sachant la base de données d.*

$$
P(\mu/d, G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \widetilde{\Pi}_{N_{ij}+t_{ij}, \frac{\sum_{h \in M_{ij}} x_i^{(h)} + t_{ij} m_{ij}}{N_{ij} + t_{ij}}}.
$$

Preuve 3.3. *En appliquant la formule de Bayes, on a*

$$P(\mu/d,G) = \frac{P(d/\mu,G)P(\mu)}{P(d/G)}.$$

Ainsi,

$$P(\mu/d,G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{e^{\langle \psi(\mu_{ij}),\ \sum\limits_{h\in M_{ij}} x_i^{(h)}\rangle - N_{ij}k(\psi(\mu_{ij}))}\widetilde{K}_{t_{ij},m_{ij}}e^{t_{ij}\langle\psi(\mu_{ij}),m_{ij}\rangle - t_{ij}k(\psi(\mu_{ij}))}}{\prod\limits_{i=1}^{n}\widetilde{K}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}}{N_{ij}+t_{ij}}}},$$

après simplification, on obtient

$$\begin{aligned} P(\mu/d,G) &= \prod_{i=1}^{n}\prod_{j=1}^{q_i} e^{\langle \psi(\mu_{ij}),\ \sum\limits_{h\in M_{ij}} x_i^{(h)} + t_{ij}m_{ij}\rangle - (N_{ij}+t_{ij})k(\psi(\mu_{ij}))} \\ &\quad \widetilde{K}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}}{N_{ij}+t_{ij}}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{q_i} \widetilde{\Pi}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}}{N_{ij}+t_{ij}}} = \prod_{i=1}^{n}\prod_{j=1}^{q_i} P(\mu_{ij}/d,G). \end{aligned}$$

L'estimateur EAP de μ_{ij} est donné par

$$\begin{aligned} \widehat{\mu}_{ij}^{EAP} &= \int \mu_{ij}P(\mu_{ij}/d,G)d\mu_{ij} \\ &= \int \mu_{ij}\widetilde{\Pi}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}}{N_{ij}+t_{ij}}} d\mu_{ij}. \end{aligned}$$

Pour calculer cet estimateur, on utilise la méthode de Monte-Carlo.
On note que l'estimateur MAP de μ_{ij} est donné par

$$\widehat{\mu}_{ij}^{MAP} = \frac{\sum\limits_{h\in M_{ij}} x_i^{(h)} + t_{ij}m_{ij}}{N_{ij}+t_{ij}}.$$

3.2.5 Exemples

Nous illustrons nos résultats en donnant les exemples suivants. Nous appliquons les résultats précédents au réseau Bayésien exponentiel discret dans des modèles particuliers comme Poisson et Binomial Négatif.

• Modèle de Poisson : La constante de normalisation est

$$\widetilde{K}_{(t_{ij},m_{ij})} = \frac{t_{ij}^{t_{ij}m_{ij}+1}}{\Gamma(t_{ij}m_{ij}+1)}.$$

La fonction score est

$$gBD(d,G) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{t_{ij}^{t_{ij}m_{ij}+1}}{(N_{ij}+t_{ij})^{\sum_{h\in M_{ij}} x_i^{(h)}+t_{ij}m_{ij}+1}}\frac{\Gamma(t_{ij}m_{ij}+\sum_{h\in M_{ij}} x_i^{(h)}+1)}{\Gamma(t_{ij}m_{ij}+1)}.$$

• Modèle Binomial Négatif : On a

$$\widetilde{K}_{t_{ij},m_{ij}} = \frac{\Gamma(t_{ij}m_{ij}+t_{ij})}{\Gamma(t_{ij}m_{ij}+1)\Gamma(t_{ij}-1)},\ t_{ij}>1.$$

Avec cette constante de normalisation, on obtient la fonction score proposée

$$gBD(d,G) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(t_{ij}m_{ij}+t_{ij})\Gamma(t_{ij}m_{ij}+\sum_{h\in M_{ij}} x_i^{(h)}+1)\Gamma(N_{ij}+t_{ij}-1)}{\Gamma(t_{ij}m_{ij}+1)\Gamma(t_{ij}m_{ij}+\sum_{h\in M_{ij}} x_i^{(h)})\Gamma(t_{ij}-1)}.$$

3.3 Travaux apparentés

Comme initialement proposé par [23], nous nous intéressons à l'extension de la distribution des variables à la famille exponentielle naturelle (FEN), qui représente une classe de distributions très importante en probabilités et statistique. Cette idée a déjà été développé par Beal et Ghahramani [3] (modèles exponentiels conjugué) pour les réseaux bayésiens avec variables latentes. Ils ont concentré leur travail sur la maximisation Expectation (EM) de l'estimation nécessaire en raison de variables latentes, mais ils n'ont pas précisé les estimateurs bayésiens utilisés, ce qui limite donc leurs expériences pour les distributions multinomiales habituels. Wainwright et Jordan [67] ont également proposé une étude intéressante sur les modèles graphiques comme familles exponentielles, montrant que les structures très spécifiques de modèles graphiques probabilistes dirigés ou non dirigés peuvent être interprétés comme des distributions exponentielles. Notre travail poursuit avec la même idée générale, traitant des RBs exponentiels discrets au lieu des RBs usuels pour explorer une gamme élargie de modèles probabilistes.

3.4 Expérimentations

3.4.1 Les données

Ilustrons l'intérêt d'utiliser les RBeds au lieu des RBs usuels pour l'estimation de la densité, en effectuant des simulations dans plusieurs contextes.

Dans le premier contexte, les données sont générées à partir des distributions décrites par les RBs usuels (dist=multi). Dans le second contexte, les données sont générées à partir des distributions décrites par les RBeds Poisson (dist=Poisson).
Dans ces contextes, nous pouvons contrôler plusieurs paramètres tels que le nombre n de variables ($n = 10, 30, 50$) et la taille M de l'ensemble des données générées ($M = 100, 1000, 10000$). La cardinalité maximale K de nos variables discrètes est également contrôlée par les RBs usuels ($K = 2, 3, 5$), mais mesurée dans les échantillons générés pour les RBeds Poisson.
Chaque génération d'ensemble de données dans de telles conditions est itéré 10×10 fois, avec 10 DAG générés aléatoirement, et 10 valeurs des paramètres aléatoires pour chacun de ces DAG.

3.4.2 Modèles et algorithmes utilisés

Notre objectif est de comparer les performances des modèles RB discrets usuels (modèle=multi) par rapport au Poisson RBed (modèle=Poisson) estimés avec les ensembles de données antérieures.
Les paramètres a priori sont choisis dans leur forme la plus simple, $\alpha_{ij} = 1$, coefficient uniforme de Dirichlet pour les RBs discrets et $t_{ij} = 1, m_{ij} = 0$ pour RBeds Poisson.
La procédure d'apprentissage de structure utilisée pour optimiser la fonction score Bayésien est la recherche gloutonne utilisée dans [9].
Afin d'obtenir des résultats plus robustes, cette recherche gloutonne est réalisée 10 fois avec des initialisations aléatoires différentes et le meilleur résultat des 10 pistes est maintenu.
Pour l'apprentissage des paramètres on utilise l'estimation du maximum a posteriori.
Nos différents modèles et algorithmes ont été implémenté sous Matlab avec BNT [54] et la "toolbox" apprentissage de structure [46].

3.4.3 Les critères d'évaluation

L'évaluation de la précision de chaque modèle est estimée par la divergence de Kullback-Leibler (KL) divergence entre la distribution d'origine décrite

par le modèle utilisé pour générer un ensemble de données et la distribution finale obtenue par le modèle estimé de cette base de données. Pour un grand nombre de configurations des variables (plus de 10^5), une approximation MCMC est utilisée avec 10^5 configurations aléatoires.
La comparaison des deux modèles est illustrée en traçant les valeurs absolues de KL obtenues par les RBeds contre les RBs usuels pour les mêmes ensembles de données.
Le fait qu'un modèle est meilleur que l'autre peut être observé à l'égard de la première diagonale (triangle supérieur : RBed est meilleur, par rapport au triangle inférieur : RB usuel est meilleur). Afin de déterminer si les différences sont statistiquement significative, nous utilisons le test de rangs signés de Wilcoxon pour des échantillons appariés avec un niveau de signification égal à 0.05 pour les 100 expériences effectuées dans un contexte donné (dist, $n; M; K$).

3.4.4 Résultats et Interprétations

Nos résultats sont décrits dans la figure 3.1 concernant $n = 10$ mais un comportement similaire est obtenu pour $n = 30$. Comme on peut le voir, lorsque les données sont générées à partir des distributions de Poisson (résultats en magenta), nos RBeds Poisson sont des modèles meilleurs que les RBs usuels. Lorsque les données sont générées à partir des distributions multinomiales (résultats en bleu), les résultats dépendent de la taille de l'échantillon M. Lorsque M est élevé ($M = 10000$, troisième figure à droite), RBs usuels sont de meilleurs modèles que RBeds Poisson. Quand la taille de l'échantillon diminue ($M = 1000$), les RBs usuels et les RBeds donnent des résultats similaires. Avec un échantillon de petite taille ($M = 100$), les RBeds sont meilleurs que les RBs usuels.
Tous ces résultats sont confirmés par des tests de Wilcoxon. En comparant les résultats à l'égard de la variable de cardinalité maximale K nous pouvons observer que les RBeds et les RBs donnent des résultats similaires pour $K = 3$, mais la situation change si K augmente. Pour $K = 6$, les RBeds donnent de meilleurs résultats que les RBs. Ces résultats sont présentés dans la figure 3.2.

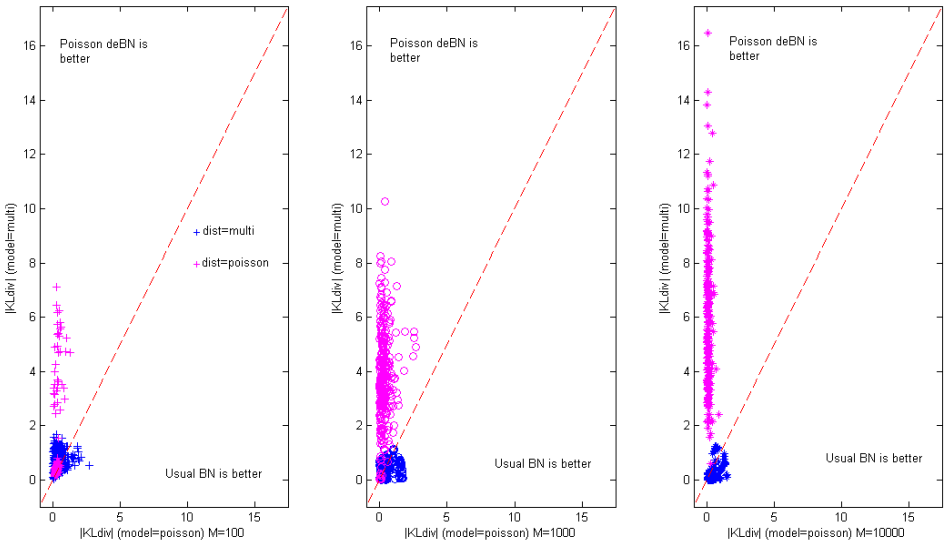

FIG. 3.1 – Comparaison de la divergence KL obtenue par les RBeds Poisson contre les RBs usuels pour les mêmes bases de données (triangle supérieur : RBed est meilleur, par rapport triangle inférieur : RB usuel est mieux) en ce qui concerne la taille des données ($M = 100, 1000, 10000$) et de la distribution de ces données (*dist=Poisson ou multinomiale*) pour $n = 10$.

On peut observer que les RBeds semblent être une bonne alternative aux RBs usuels dans plusieurs contextes. Si on compare les RBeds Poisson et les RBs usuels, les premiers ont moins de paramètres libres que les autres et ce nombre de paramètres est moins dépendant de la valeur de K. Ainsi, lorsque la taille M de l'échantillon est faible ou lorsque la valeur de K est élevée, les RBeds sont un bon compromis pour l'estimation de la densité.

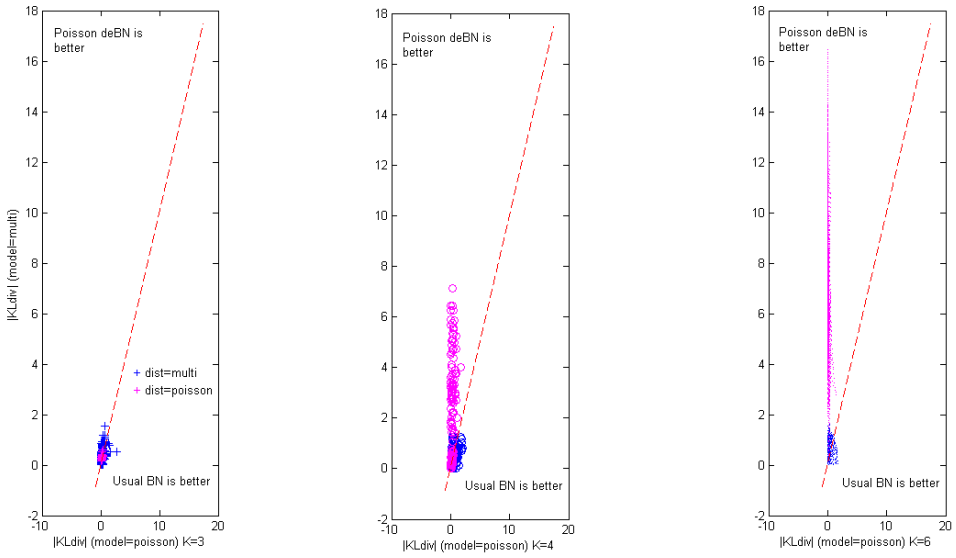

FIG. 3.2 – Comparaison de la divergence KL obtenue par les RBeds contre les RBs usuels pour les mêmes ensembles de données (triangle supérieur : RBed est meilleur, par rapport au triangle inférieur : RB usuel est mieux) par rapport à la cardinalité maximale (K = 3, 4, 6) et de distribution des données (*dist* = *Poisson ou multinomiale*) pour $n = 10$ et $n = 30$.

3.5 Conclusion

Dans ce chapitre, nous avons développé le concept de réseau Bayésien exponentiel discret.

Grâce à l'utilisation de la famille à priori $\widetilde{\Pi}$ qui généralise pour toute distribution de FEN la distribution de Dirichlet utilisé pour les réseaux Bayésiens discrets, nous avons proposé une nouvelle fonction de score qui étend le score Dirichlet Bayésien. Nous proposons également des estimateurs Bayésiens pour les paramètres de réseau Bayésien exponentiel discret qui peuvent être utilisés pour n'importe quelle distribution de FEN.

Chapitre 4

Estimation implicite des paramètres dans des modèles Gaussiens

4.1 Introduction

Dans ce chapitre, nous étudions le problème d'apprentissage des paramètres pour trois modèles Gaussiens en utilisant les deux approches implicite et Bayésienne.
D'abord, on s'intéresse à la Gaussienne simple puis à un modèle à volatilité stochastique, le modèle de Heston. Et finalement, on va étudier les réseaux Bayésiens Gaussiens conditionnels. On reprend les résultats du travail de Murphy [55] décrivant l'estimation des paramètres pour les modèles Gaussiens conditionnels en les élargissant avec l'approche implicite et en comparant les estimateurs obtenus par les deux approches.

4.2 Estimation de la variance d'une loi Gaussienne

Soit $X_i \sim N(0, \sigma^2)$ une variable aléatoire Gaussienne centrée de variance connue $\sigma^2 \in]0, +\infty[$, la vraisemblance de σ^2 pour n observations indépen-

dantes $\underline{x} = (x_1, \ldots, x_n)$ est

$$l(\underline{x}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\}.$$

Par un calcul standard, on remarque que l'estimateur implicite de σ^2 est

$$(\widehat{\sigma}^2)^{Imp} = E(\sigma^2 | X_1, \ldots, X_n) = \frac{1}{n-4} \sum_{i=1}^{n} X_i^2, \qquad n > 4 \qquad (4.1)$$

qui est différent de l'estimateur du Maximum de Vraisemblance dont le facteur de normalisation est $\frac{1}{n}$. On peut comparer ce résultat avec l'estimation Bayésienne obtenue par l'approche d'espérance a posteriori (EAP)

$$(\widehat{\sigma}^2)^{Bay} = \frac{\sum_{i=1}^{n} X_i^2 + 2b}{n + 2a - 2}, \qquad (4.2)$$

où la loi a priori de σ^2 est une Inverse Gamma $IG(a, b)$ où a est un paramètre de forme et b est un paramètre d'échelle.

Nous constatons aussi que l'estimateur Bayésien de la variance est égal à l'estimateur implicite pour $a = -1$ et $b = 0$, qui n'est pas possible car a et b, sont les paramètres d'une distribution Inverse Gamma, doivent être positifs. Mukhopadhyay [52] a affirmé que l'inférence implicite n'est pas nouvelle et qu'elle est soit une approche Fudicielle ou une méthode Bayésienne non informative. Concernant ses critiques, notre commentaire est un nouveau paradigme dans l'inférence statistique. En utilisant plusieurs exemples, nous montrons que, dans de nombreux cas lorsque l'espace des paramètres θ est non borné, la distribution implicite ne coïncide pas ni avec la distribution Bayésienne ni avec la distribution Fiducielle. Si l'espace des paramètres θ est borné, alors la distribution implicite coïncide avec la distribution a posteriori avec a priori uniforme dans la méthode Bayésienne. La coïncidence de ces deux distributions implicite et Fiducielle pour la loi Normale $N(\theta, 1)$ de moyenne θ et de variance 1 explique les trompeuses commentaires de [52].

On s'intéresse à comparer les deux approches Bayésienne et implicite à travers l'estimateur de σ^2.

Pour évaluer la performance de la méthode implicite, on procède dans une première étape par des simulations de données en utilisant le logiciel Matlab,

dans une deuxième étape nous validerons nos résultats en les comparant avec la méthode Bayésienne.
Pour comparer une approche statistique à une autre et pour apprécier dans quelle mesure le résultat sera plus précis, on opte pour un indicateur qui est le carré moyen des erreurs ou erreur moyenne quadratique (MSE pour *Mean Squared Error*). Comme son nom l'indique c'est la moyenne arithmétique des carrés des écarts entre l'estimateur et la valeur vraie du paramètre.
On génère 1000 observations du modèle Gaussien et on compare l'estimateur Bayésien et implicite du point de vue MSE pour des différentes valeurs vraies du paramètre σ^2. Les résultats sont présentés dans le tableau suivant

$\overline{\sigma^2}$	a priori	$(\overline{\widehat{\sigma^2}})^{Imp}$	MSE(Imp)	$(\overline{\widehat{\sigma^2}})^{Bay^*}$	$MSE(Bay^*)$	$(\overline{\widehat{\sigma^2}})^{Bay^{**}}$	$MSE(Bay^{**})$
0.1	a=2, b=0.1	0.1	$2\ .\ 10^{-6}$	0.1	$2\ .\ 10^{-6}$	0.11	$1.03.\ 10^{-5}$
0.05	a=2, b=0.05	0.05	$4.99.\ 10^{-7}$	0.05	$5.\ 10^{-7}$	0.053	$9.39.\ 10^{-6}$
1	a=2, b=1	1.0002	$1.99.\ 10^{-4}$	1.0006	$1.99.\ 10^{-4}$	1.019	$5.87.\ 10^{-4}$
1.5	a=2, b=1.5	1.5	$4.45.\ 10^{-4}$	1.5006	$4.46.\ 10^{-4}$	1.52	0.0013
2	a=2, b=2	2	$8.03.\ 10^{-4}$	2.0007	$8.05.\ 10^{-4}$	2.039	0.0024

TAB. 4.1 – Estimateurs de σ^2 obtenus par les équations (4.1) et (4.2)

(*) Les paramètres estimés obtenus en utilisant les valeurs vraies comme des paramètres a priori.
(**) Les paramètres estimés obtenus en utilisant des données a priori différentes des valeurs vraies.

4.2.1 Analyse des résultats

En comparant l'approche implicite à celle Bayésienne basée sur les valeurs vraies (valeurs avec lesquelles nous avons simulé les données, prises comme données a priori et qui sont les probabilités les plus favorables pour l'estimation Bayésienne), on constate une concordance excellente entre les deux approches manifestées par des précisions très similaires pour les paramètres estimés. Par contre, nous remarquons une meilleure précision des paramètres estimés par la méthode implicite comparés à ceux obtenus par la méthode Bayésienne en utilisant des données a priori différentes des valeurs vraies. Ce

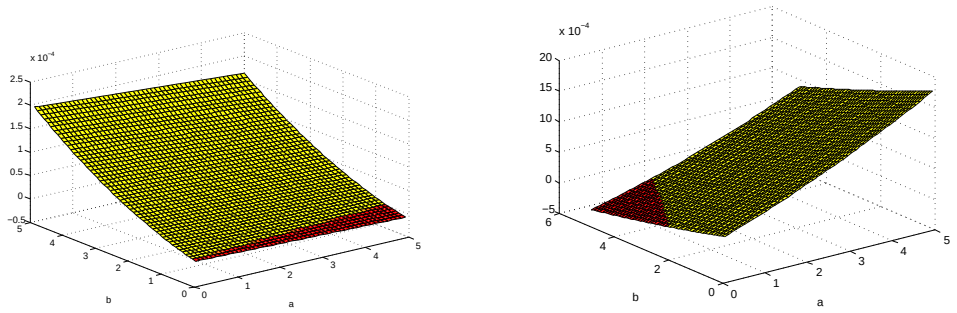

FIG. 4.1 – La différence entre l'erreur moyenne quadratique Bayésienne et implicite pour $\sigma^2 = 0.1$ (gauche) et $\sigma^2 = 2$ (droite) pour différentes valeurs des paramètres a priori a et b.

résultat prouve la fragilité de l'estimation Bayésienne : Des mauvais résultats peuvent être obtenus si nous fixons un mauvais choix pour la distribution a priori.
Pour illustrer ces résultats, la Figure 4.1 présente la différence entre MSE Bayésien et MSE implicite pour différentes valeurs de σ^2 en faisant varier les valeurs des paramètres a priori de a et b. La partie jaune correspond à la région où le MSE implicite est plus petit que celui Bayésien. Quel que soit la valeur de σ^2 (petite ou grande), MSE implicite est dans la majorité plus bas que le MSE Bayésien.

4.3 Estimation des paramètres dans les réseaux Bayésiens associés à un modèle de Heston

Après avoir présenter le principe des approches Bayésienne et implicite, nous allons l'appliquer pour l'estimation d'un modèle à volatilité stochastique.

4.3.1 Présentation du modèle de Heston

Ce modèle, introduit par Heston [30], représente une généralisation du modèle Black et Scholes [7] du fait qu'il incorpore une volatilité qui varie avec le temps avec bruit supplémentaire lors du calcul du prix de l'option. Ce modèle fournit une solution analytique pour calculer le prix d'une option d'achat lorsqu'il y a une corrélation entre le prix du sous-jacent et sa vola-

tilité. Il peut être adapté pour incorporer des taux d'intérêts stochastiques et par conséquent il peut être utilisé pour calculer le prix des options sur obligations et devises étrangères.
Le processus de diffusion qui décrit le prix de l'option est identique à celui de Black et Scholes, à l'exception de la volatilité qui varie dans le temps. La volatilité suit le processus racine carré utilisé par Cox, Ingersoll et Ross [17]. Dans ce modèle les dynamiques des prix et de la volatilité sont données par ces équations différentielles stochastiques suivantes :

$$\begin{cases} dS_t &= \mu_t S_t dt + \sqrt{V_t} S_t dW_t^s \\ dV_t &= k(\theta - V_t)dt + \sigma\sqrt{V_t} dW_t^v, \end{cases} \tag{4.3}$$

où W_t^s et W_t^v sont deux mouvements browniens corrélés avec $\rho = corr(W_t^s, W_t^v)$,
• μ_t : taux de rendement espéré instantané de l'action (drift parameter),
• $\sqrt{V_t}$: volatilité du rendement,
• θ : la moyenne de la variance à long terme,
• k : un paramètre de retour à la moyenne,
• σ : l'écart type de la volatilité.
L'équation (4.3) a une solution positive si $2k\theta > \sigma^2$ et $V_0 > 0$ c'est la condition de Feller [1].
Il est important de signaler que la façon la plus naturelle pour simuler ce processus est la discrétisation d'Euler. En prenant $\mu_t = 0$, on obtient le système d'équations suivant

$$\begin{cases} Y_t &= \sqrt{V_t}\epsilon_t^s \\ V_{t+1} &= k\theta + (1-k)V_t + \sigma\sqrt{V_t}\ \epsilon_t^v \end{cases} \tag{4.4}$$

avec $(\epsilon_t^s, \epsilon_t^v)$ est un vecteur gaussien centré et corrélé. Notons $\rho = Corr(\epsilon_t^s, \epsilon_t^v)$ le coefficient de corrélation linéaire de ϵ_t^s et ϵ_t^v.
Il est important de noter aussi qu'il y a un problème de discrétisation pour le modèle de Heston. Comme l'erreur dans l'équation (4.4) est supposée normale, alors la volatilité V_t peut être négative avec une probabilité non nulle. Pour résoudre ce problème de discrétisation, Deelstra et Delbaen [19] ont proposé le système d'équations suivant

$$\begin{cases} Y_t &= \sqrt{V_t}\epsilon_t^s \\ V_{t+1} &= k\theta + (1-k)V_t + \sigma\sqrt{V_t \mathbf{1}_{V_t>0}}\ \epsilon_t^v. \end{cases}$$

On note $\alpha = k\theta$, $\beta = 1 - k$. On définit $V = (V_t)_{t=1}^T$ et $\Theta = (\alpha, \beta, \sigma^2)$ le vecteur de paramètres.

4.3.2 Estimation des paramètres par la méthode Bayésienne

On suppose que les a prioris pour les paramètres α et β sont des lois normales centrées réduites (N), quant à σ^2 suit une loi inverse gamma (IG). Comme les a posterioris sont conjuguées, alors $P(\alpha, \beta/\sigma, V, Y)$ est normale et $P(\sigma^2/\alpha, \beta, V, Y)$ est une inverse gamma. Ceci implique que les densités a posteriori des paramètres sont proportionnelles à

$$P(\alpha, \beta/\sigma, V, Y) \propto \prod_{t=1}^{T} P(V_t/V_{t-1}, \alpha, \beta, \sigma) P(\alpha, \beta) \propto N,$$

pour σ^2, on a

$$P(\sigma^2/\alpha, \beta, V, Y) \propto \prod_{t=1}^{T} P(V_t/V_{t-1}, \alpha, \beta, \sigma) P(\sigma^2) \propto IG.$$

A l'itération j, en appliquant l'approche Bayésienne, les estimateurs de α, β et σ^2 sont définies par

$$\alpha^{(j)} = \frac{\sum_{t=1}^{T} \left(\frac{V_t^{(j-1)}}{V_{t-1}^{(j-1)}} - \beta^{(j-1)} \right)}{(\sigma^2)^{(j-1)} + \sum_{t=1}^{T} \frac{1}{V_{t-1}^{(j-1)}}},$$

$$\beta^{(j)} = \frac{\sum_{t=1}^{T} \left(V_t^{(j-1)} - \alpha^{(j)} \right)}{(\sigma^2)^{(j-1)} + \sum_{t=1}^{T} V_{t-1}^{(j-1)}}$$

et

$$(\sigma^2)^{(j)} = \frac{b + \frac{1}{2} \sum_{t=1}^{T} \left(\frac{(V_t^{(j-1)} - \alpha^{(j)} - \beta^{(j)} V_{t-1}^{(j-1)})^2}{V_{t-1}^{(j-1)}} \right)}{a + \frac{T}{2} - 1},$$

avec a et b sont les paramètres de la densité a priori d'une loi inverse gamma pour σ^2 [25].

4.3.3 Estimation des paramètres par la méthode implicite

La fonction vraisemblance est définie par

$$\prod_{t=1}^{T} P(V_t/V_{t-1},\alpha,\beta,\sigma) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2 V_{t-1}}}\; e^{-\frac{(V_t-\alpha-\beta V_{t-1})^2}{2\sigma^2 V_{t-1}}}.$$

La loi implicite de α est une normale

$$N\Big(\frac{\sum_{t=1}^{T}\frac{V_t}{V_{t-1}} - \beta T}{\sum_{t=1}^{T}\frac{1}{V_{t-1}}}, \frac{\sigma^2}{\sum_{t=1}^{T}\frac{1}{V_{t-1}}}\Big).$$

A une itération (j), l'estimateur implicite de α est

$$\alpha^{(j)} = \frac{\sum_{t=1}^{T}\frac{V_t^{(j-1)}}{V_{t-1}^{(j-1)}} - \beta^{(j-1)} T}{\sum_{t=1}^{T}\frac{1}{V_{t-1}^{(j-1)}}}.$$

La loi implicite de β est aussi une normale

$$N\Big(\frac{\sum_{t=1}^{T}(V_t-\alpha)}{\sum_{t=1}^{T}V_{t-1}}, \frac{\sigma^2}{\sum_{t=1}^{T}V_{t-1}}\Big).$$

Donc, l'estimateur implicite de β, à une itération (j), est égal à

$$\beta^{(j)} = \frac{\sum_{t=1}^{T}\Big(V_t^{(j-1)} - \alpha^{(j)}\Big)}{\sum_{t=1}^{T}V_{t-1}^{(j-1)}}.$$

La densité de la loi implicite de σ^2 est une inverse gamma

$$\frac{\sigma^{-T}\exp\Big(-\frac{\sum_{t=1}^{T}(\frac{V_t-\alpha-\beta V_{t-1}}{\sqrt{V_{t-1}}})^2}{2\sigma^2}\Big)}{\Gamma(\frac{T}{2}-1)}\Big(\frac{1}{2}\sum_{t=1}^{T}(\frac{V_t-\alpha-\beta V_{t-1}}{\sqrt{V_{t-1}}})^2\Big)^{\frac{T}{2}-1}.$$

A une itération (j), l'estimateur implicite de σ^2 est

$$(\sigma^2)^{(j)} = \frac{1}{T-4}\sum_{t=1}^{T}\Big(\frac{V_t^{(j-1)} - \alpha^{(j)} - \beta^{(j)}V_{t-1}^{(j-1)}}{\sqrt{V_{t-1}^{(j-1)}}}\Big)^2.$$

4.3.4 Comparaison

Dans cette section, on s'intéresse à comparer les deux approches Bayésienne et implicite à travers le modèle de Heston.
Nous allons présenter les résultats numériques d'estimation du modèle. Pour ce faire, nous allons mener une analyse de simulation de données. Le critère de l'erreur moyenne quadratique nous permet d'évaluer l'approche Bayésienne par rapport à l'approche implicite.
Le tableau suivant résume l'étude de simulation que nous avons effectué. On donne la valeur vraie du paramètre, ses estimateurs Bayésiens (Bay (*), Bay (**)), celui implicite. On donne, en plus, l'erreur moyenne quadratique calculée pour chaque estimateur [25].

paramètres	α	β	σ
valeurs vraies	$0,33$	$0,3$	$0,04$
Estimation Bayésienne (*)	$0,209$	$0,55$	$0,05$
MSE Bayésien (*)	$4,31.10^{-5}$	$1,95.10^{-4}$	$2,8.10^{-4}$
Estimation implicite	$0,216$	$0,56$	$0,05$
MSE implicite	$1,33.10^{-5}$	$1,99.10^{-4}$	$2,7.10^{-4}$
Estimation Bayésienne (**)	$0,19$	$0,65$	$0,07$
MSE Bayésien (**)	$2,19.10^{-4}$	$9,95.10^{-4}$	$4,2.10^{-4}$

TAB. 4.2 – Simulation des données pour le modèle de Heston en utilisant l'approche implicite, Bayésienne (*) (Les paramètres estimés obtenus en utilisant les valeurs vraies comme des paramètres à priori) et Bayésienne (**) (Les paramètres estimés obtenus en utilisant des données a priori différentes des valeurs vraies).

La comparaison de MSE obtenu par Bay (*) et Bay(**) prouve la sensibilité de l'estimation des paramètres selon le choix de la distribution a priori. Ce résultat prouve la fragilité de l'estimation Bayésienne. Notre estimation implicite donne des résultats plus robustes que ceux obtenus par la méthode Bayésienne, en particulier si les a prioris utilisés dans l'estimation Bayésienne sont loin des valeurs vraies.
En conclusion, grâce au critère de l'erreur quadratique moyenne, et en se

basant sur une base de données simulée, on a trouvé que les estimateurs issus de la méthode implicite sont les plus proches des valeurs vraies. Ceci prouve que cette approche est meilleure que l'approche Bayésienne du point de vue de la précision des estimateurs.

4.4 Estimation des paramètres dans les réseaux Bayésiens Gaussiens conditionnels

4.4.1 Notations

Lorsque les variables dans un réseau ont des densités linéaire Gaussienne conditionnelle, la densité jointe de X est une distribution Gaussienne multivariée.
Pour simplifier les équations qui suivent, on utilise les mêmes notations proposées par Murphy [55] utilisées dans la section 1.5.2. On considère le problème de trouver les estimateurs des paramètres d'une variable Gaussienne conditionnelle Y avec des parents continus X et des parents discrets Q, *i.e*

$$p(y|x, Q = i) = c|\Sigma_i|^{-\frac{1}{2}} exp(-\frac{1}{2}(y - B_i x - \mu_i)'\Sigma_i^{-1}(y - B_i x - \mu_i))$$

où x et y sont des réalisations des variables aléatoires X et Y respectivement, c est une constante et $|y| = d$. Considérons un N-échantillon aléatoire, la fonction vraisemblance est définie par

$$\prod_{t=1}^{N}\prod_{i=1}^{|Q|} p(y_t|x_t, Q_t = i, e_t)^{q_t^i}$$

où $q_t^i = 1$ si Q a la valeur i dans la t^{eme} configuration et 0 sinon. Lorsque Q, X et Y ne sont pas observées, on définit la moyenne de la vraisemblance. Soit $\widetilde{p}(y|x, Q = i) = \exp(l)$ avec

$$l = -\frac{1}{2}\sum_{t=1}^{N} E\Big(\sum_{i=1}^{|Q|} q_t^i \log|\Sigma_i| + q_t^i(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)|e_t\Big).$$

On suppose que

$$E(q_t^i x_t x_t'|e_t) = E(q_t^i|e_t)E(x_t x_t'|Q_t = i, e_t) = w_t^i E_{ti}(XX')$$

où les poids $w_t^i = p(Q = i/e_t)$ sont les probabilités a posteriori et $E_{ti}(XX')$ est le deuxième moment conditionnel. On obtient

$$l = -\frac{1}{2}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i \log|\Sigma_i| - \frac{1}{2}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i E_{ti}\Big((y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)|e_t\Big).$$

$$w_i = \sum_{t=1}^{N} w_t^i,\ S_{XX',i} = \sum_{t=1}^{N} w_t^i E_{ti}(XX') \text{ et } S_{X,i} = \sum_{t=1}^{N} w_t^i E_{ti}(X).$$

Dans cette section, nous inspirons du travail de Murphy consacré à l'estimation des paramètres de modèles Gaussiens conditionnels par la méthode du maximum à posteriori (MAP) et par maximum de vraisemblance (MV) en les élargissant avec l'estimation d'espérance a posteriori (EAP) et l'estimation implicite.

4.4.2 Estimation bayésienne des paramètres

Estimation de la matrice de régression B_i

Lemme 4.1. *On suppose que l'a priori de B_i est une distribution Normale multivariée de moyenne a et de matrice de covariance V. Ainsi, l'estimateur de B_i est égale à*

$$\widehat{B_i}^{Bay} = \Big(\Sigma_i^{-1}(S_{YX',i} - \mu_i S_{X',i}) + aV^{-1}\Big)\ \Big(\Sigma_i^{-1} S_{XX',i} + V^{-1}\Big)^{-1}. \quad (4.5)$$

Preuve 4.1. *La vraisemblance est définie par*

$$\widetilde{p}(y|B_i, x_t, \Sigma_i, \mu_i) \propto \exp(-\frac{1}{2}\sum_{t=1}^{N}(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)).$$

L'a posteriori $\widetilde{p}(B_i|y_t, x_t, \Sigma_i, \mu_i)$ est une distribution Normale multivariée.

$$\widetilde{p}(B_i/y_t, x_t, \Sigma_i, \mu_i) \propto \exp(-\frac{1}{2}\sum_{t=1}^{N}(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)) \exp(-\frac{1}{2}(B_i - a)'$$

$$V^{-1}(B_i - a)).$$

En utilisant l'égalité suivante :

$$\frac{\partial(XA + b)'C(XA + b)}{\partial X} = (C + C')(XA + b)A' \quad (4.6)$$

on trouve l'expression de l'estimateur de B_i.

Estimation de la moyenne μ_i

Lemme 4.2. *On suppose que la densité a priori pour le paramètre μ_i est une distribution Normale multivariée de paramètres (m, ψ) [58]. Donc, l'estimateur Bayésien de la moyenne est donné par cette expression*

$$\widehat{\mu_i}^{Bay} = (\psi^{-1}m + \Sigma_i^{-1}(S_{Y,i} - B_i S_{X,i}))\ (w_i\Sigma_i^{-1} + \psi^{-1})^{-1}. \tag{4.7}$$

Preuve 4.2. *La densité a posteriori de μ_i est égale à*

$$\widetilde{p}(\mu_i/y_t, x_t, \Sigma_i, B_i) \propto exp\Big(-\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{ti}(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)\Big)$$

$$exp\Big(-\frac{1}{2}(\mu_i - m)'\psi^{-1}(\mu_i - m)\Big) \tag{4.8}$$

et en utilisant l'égalité (4.6) on trouve l'expression suivante de l'estimateur de la moyenne

$$\widehat{\mu_i}^{Bay} = (\psi^{-1}m + \Sigma_i^{-1}(S_{Y,i} - B_i S_{X,i}))\ (w_i\Sigma_i^{-1} + \psi^{-1})^{-1}.$$

Estimation de la matrice de covariance Σ_i

Lemme 4.3. *On suppose que l'a priori de Σ_i suit une distribution Inverse-Wishart de degrés de liberté α et une matrice de précision définie positive V. Ces hypothèses impliquent que l'estimateur Bayésien de Σ_i est défini par*

$$\widehat{\Sigma_i}^{Bay} = \frac{1}{w_i + \alpha - d - 1}(A_i + V) \tag{4.9}$$

où $A_i = S_{YY',i} - S_{YX',i}B_i' - S_{Y,i}\mu_i' - B_i S_{XY',i} + B_i S_{XX',i}B_i' + B_i S_{X,i}\mu_i' - \mu_i S_{Y',i} + \mu_i S_{X',i}B_i' + \mu_i\mu_i'$.

Preuve 4.3. *La densité a posteriori du paramètre Σ_i est proportionnelle à*

$$\begin{aligned}\widetilde{p}(\Sigma_i|y_t, x_t, B_i, \mu_i) \quad \propto \quad & |\Sigma_i|^{-\frac{w_i+\alpha+d+1}{2}} \exp(-\tfrac{1}{2}tr(\Sigma_i^{-1}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i E_{ti}(y_t - B_i x_t - \mu_i) \\ & (y_t - B_i x_t - \mu_i)' + V).\end{aligned}$$

Donc, l'a posteriori $\widetilde{p}(\Sigma_i|B_i, y_t, x_t, \mu_i)$ est aussi une distribution Inverse-Wishart de degrés de liberté $w_i+\alpha$ et une matrice de précision définie positive

$$\sum_{t=1}^{N} w_t^i E_{ti}(y_t - B_i x_t - \mu_i)(y_t - B_i x_t - \mu_i)' + V.$$

Ceci implique que

$$\widehat{\Sigma_i}^{Bay} = \frac{1}{w_i + \alpha - d - 1}(A_i + V).$$

Estimation d'une matrice de covariance avec $\Sigma_i = \sigma_i^2 I$

Lemme 4.4. *On suppose que la distribution a priori pour σ_i^2 est une Inverse Gamma de paramètres (a, b). Donc, l'estimateur de σ_i^2 est égale à*

$$(\widehat{\sigma_i^2})^{Bay} = \frac{1}{dw_i + 2a - 2} tr(C_i + 2b) \tag{4.10}$$

où $C_i = S_{Y'Y,i} - B_i S_{Y'X,i} - S_{Y',i}\mu_i - B_i' S_{X'Y,i} + B_i' B_i S_{X'X,i} + B_i'\mu_i S_{X',i} - \mu_i' S_{Y,i} + \mu_i' B_i S_{X,i} + \mu_i'\mu_i$.

Preuve 4.4. *Si on a la contrainte que $\Sigma_i = \sigma_i^2 I$, la densité conditionnelle de Y est*

$$p(y|x, Q = i) = c\sigma_i^{-d} exp(-\frac{1}{2}\sigma_i^{-2}\|y - B_i x_t - \mu_i\|^2)$$

et en supposant que la distribution a priori pour σ_i^2 est une Inverse Gamma de paramètres (a, b)

$$\widetilde{p}(\sigma_i^2|y_t, x_t, B_i, \mu_i) \propto (\sigma_i^2)^{-\frac{dw_i}{2} - a - 1} \exp\Big(-\frac{1}{\sigma_i^2}(\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{ti}(\|y_t - B_i x_t - \mu_i\|^2) + b)\Big).$$

Or

$$\|y_t - B_i x_t - \mu_i\|^2 = (y_t - B_i x_t - \mu_i)'(y_t - B_i x_t - \mu_i)$$

et

$$x'Ay = tr(x'Ay) = tr(Ayx')$$

en utilisant le fait que

$$E(x'Ay) = tr(AE(yx'))$$

on trouve l'expression suivante de l'estimateur de σ_i^2

$$(\widehat{\sigma_i^2})^{Bay} = \frac{1}{dw_i + 2a - 2} tr(C_i + 2b).$$

4.4.3 Estimation implicite des paramètres

Estimation de la matrice de régression

Lemme 4.5. *L'estimateur implicite de la matrice de régression est*

$$\widehat{B_i}^{Imp} = (S_{YX',i} - \mu_i S_{X',i})(S_{XX',i})^{-1}. \tag{4.11}$$

Preuve 4.5. *La loi implicite de B_i est définie par*

$$\widetilde{p}(B_i|y_t, x_t, \Sigma_i, \mu_i) \propto \exp(-\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{ti}(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)).$$

En utilisant l'égalité (4.6) avec $A = -x_t, X = B_i$ et $b = y_t - \mu_i$.
Nous montrons que l'estimateur implicite de B_i est

$$\widehat{B_i}^{Imp} = (S_{YX',i} - \mu_i S_{X',i})(S_{XX',i})^{-1}.$$

Estimation de la moyenne

Lemme 4.6. *L'estimateur implicite de la moyenne est donné par*

$$\widehat{\mu_i}^{Imp} = \frac{S_{Y,i} - B_i S_{X,i}}{\sum_{t=1}^{N} w_t^i}. \tag{4.12}$$

Preuve 4.6. *Dans le contexte de l'estimation implicite*

$$\widetilde{p}(\mu_i|y_t, x_t, \Sigma_i, B_i) \propto \exp(-\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{ti}(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)).$$

A l'aide de l'égalité (4.6) et en prenant $X = \mu_i, A = -1$ et $b = y_t - B_i x_t$ on trouve l'estimateur implicite de la moyenne

$$\widehat{\mu_i}^{Imp} = \frac{S_{Y,i} - B_i S_{X,i}}{\sum_{t=1}^{N} w_t^i}.$$

Estimation de la matrice de covariance

Lemme 4.7. *L'estimateur implicite de Σ_i est*

$$\widehat{\Sigma_i}^{Imp} = \frac{1}{w_i - 2d - 2}(A_i). \tag{4.13}$$

Preuve 4.7. *La loi implicite de* Σ_i *est définie par*

$$\widetilde{p}(\Sigma_i|y_t,x_t,B_i,\mu_i) \propto |\Sigma_i|^{-\frac{w_i}{2}} \exp(-\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{ti}(y_t-B_ix_t-\mu_i)'\Sigma_i^{-1}(y_t-B_ix_t-\mu_i))$$

$$\propto |\Sigma_i|^{-\frac{w_i}{2}} \exp(-\frac{1}{2}tr(\Sigma_i^{-1}\sum_{t=1}^{N} w_t^i E_{ti}(y_t-B_ix_t-\mu_i)(y_t-B_ix_t-\mu_i)')$$

où $\widetilde{p}(\Sigma_i|y_t,x_t,B_i,\mu_i)$ *est une distribution Inverse-Wishart de degrés de liberté* (w_i-d-1) *et une matrice de précision définie positive*

$$\sum_{t=1}^{N} w_t^i E_{ti}(y_t-B_ix_t-\mu_i)(y_t-B_ix_t-\mu_i)'.$$

L'estimateur implicite de Σ_i *est donné par*

$$\widehat{\Sigma_i}^{Imp} = \frac{1}{w_i-2d-2}(A_i).$$

Estimation d'une matrice de covariance avec $\Sigma_i = \sigma_i^2 I$

Lemme 4.8. *Si on a la contrainte que* $\Sigma_i = \sigma_i^2 I$*, alors l'estimateur implicite de* σ_i^2 *est*

$$(\widehat{\sigma_i}^2)^{Imp} = \frac{1}{dw_i-4}tr(C_i). \tag{4.14}$$

Preuve 4.8. *La densité conditionnelle de* Y *sachant* X *et* Q *est*

$$p(y|x,Q=i) = c\sigma_i^{-d}exp(-\frac{1}{2}\sigma_i^{-2}\|y-B_ix-\mu_i\|^2)$$

$$l = -d\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i \log|\sigma_i| - \frac{1}{2}\sigma_i^{-2}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i E_{ti}\|y_t-B_ix_t-\mu_i\|^2.$$

Ceci implique que

$$\widetilde{p}(y|x,Q=i) = \exp\Big(-d\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i \log|\sigma_i| - \frac{1}{2}\sigma_i^{-2}\sum_{t=1}^{N}\sum_{i=1}^{|Q|} w_t^i E_{ti}\|y_t-B_ix_t-\mu_i\|^2\Big)$$

σ_i^2 *suit une loi Inverse Gamma où* $\frac{w_i d}{2}-1$ *est un paramètre de forme et* $\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{ti}\|y_t-B_ix_t-\mu_i\|^2$ *est un paramètre d'échelle.*
Nous déduisons donc l'estimateur implicite de σ_i^2

$$(\widehat{\sigma_i}^2)^{Imp} = \frac{\sum_{t=1}^{N} w_t^i E_{ti}(\|y_t-B_ix_t-\mu_i\|^2)}{dw_i-4}.$$

D'où

$$(\widehat{\sigma_i}^2)^{Imp} = \frac{1}{dw_i - 4} tr(C_i).$$

4.4.4 Etude comparative

Dans cette section, nous nous intéressons à comparer deux approches d'estimation à travers les réseaux Bayésiens Gaussiens conditionnels. En adoptant l'approche Bayésienne nous montrons que tous les paramètres estimés dépendent des paramètres a priori et comme nous savons, le choix de l'information a priori dans les approches Bayésiennes était toujours problématique et c'est la faiblesse importante de telles méthodes. Très souvent nous avons besoin d'un expert pour recevoir les connaissances a priori, ou l'utilisation des a prioris non informative. Ainsi le choix de la distribution a priori est une tâche problématique et si nous pouvons trouvé la distribution a posteriori avec la probabilité des données la méthode sera plus facile d'utiliser.
Comme l'approche implicite et l'approche du maximum de vraisemblance coïncident pour l'estimation du paramètre de la moyenne μ et du paramètre de régression B nous concentrons notre étude comparative sur l'estimation de la matrice de covariance. Nous remarquons que l'estimateur de $\widehat{\Sigma}$ par la méthode implicite correspond à celui obtenu par l'approche Bayésienne en prenant $V = 0$ et $\alpha = d - 1$. Ce n'est pas possible parce que V est une matrice de précision définie positive. Cette situation est semblable à celle décrite dans la section (4.2) pour l'estimateur de σ^2.
L'estimation implicite nous donne des résultats plus robustes que celle Bayésienne, en particulier si les a prioris utilisées dans l'estimation Bayésienne sont loin des valeurs vraies.

4.4.5 Travaux associés

Des premiers travaux ont montré l'intérêt de l'approche implicite pour l'apprentissage des paramètres des réseaux Bayésiens discrets dans le cadre de l'estimation des paramètres de lois multinomiales avec des données complètes [4] et incomplètes [5]. Ces travaux existants trouvent les mêmes propriétés concernant la robustesse de l'estimation implicite par rapport à l'estimation bayésienne avec des "mauvais" a prioris.

4.4.6 Validation expérimentale

Protocole expérimental

Pour valider l'approche implicite pour l'apprentissage des paramètres dans les réseaux Bayésiens Gaussiens conditionnels et mesurer la qualité de l'estimation, nous avons effectué plusieurs simulations.
Nous avons considéré différentes tailles de bases de données générées (M=100, 1000, 10000) avec un nombre de variables ($n = 10, 30, 50$). La cardinalité maximale K de ces variables est contrôlée par les réseaux Bayésiens Gaussiens conditionnels ($K = 2, 3, 5$).
Chaque ensemble de données généré dans telles conditions est itéré 10×10 fois, avec 10 DAG générés aléatoirement, et 10 valeurs des paramètres aléatoires pour chacun de ces DAG.
Notre objectif est de comparer les performances de l'approche implicite par rapport à la méthode du maximum de vraisemblance à travers les réseaux Bayésiens Gaussiens conditionnels. Nous concentrons notre étude comparative sur l'estimation de la matrice de covariance.
Pour la réalisation de ce protocole expérimental, nous avons utilisé le langage de programmation Matlab avec BNT et BNT SLP.

Mesure de la qualité de l'estimation

Pour mesurer la qualité d'estimation de chaque densité de probabilité apprise à partir des données, nous évaluons la divergence de Kullback-Leibler (KL) entre cette densité et celle qui a été utilisée pour générer les données.
Comme nous considérons un grand nombre de variables, la taille de l'ensemble de toutes les configurations possibles des variables aléatoires est très grande en fonction du nombre de variables. Donc pour un grand nombre de configurations des variables (plus de 10^5). Nous avons recours à un calcul approché de la divergence de KL en utilisant la méthode de Monte Carlo par Chaîne de Markov (MCMC) avec 10^5 configurations aléatoires.
On compare les deux méthodes d'estimation en traçant les valeurs absolues de KL obtenues par l'approche implicite et l'approche du maximum de vraisemblance pour les mêmes ensembles de données.
Le fait qu'une méthode est meilleure que l'autre est observée à l'égard de la

première diagonale (triangle supérieur : méthode de maximum de vraisemblance est meilleure, par rapport au triangle inférieur : l'approche implicite est meilleure).

Résultats et Interprétations

En comparant les résultats par rapport au nombre de variables n on observe des résultats simulaires. Donc l'étude par rapport à n n'influe pas sur les résultats, mais la situation change si on compare les résultats à l'égard de la variable de cardinalité maximale K.
Pour toutes les valeurs de la cardinalité maximale ($K = 2, 3, 5$) l'estimation Implicite est bonne. Cette dernière approche coïncide avec la méthode du maximum de vraisemblance lorsqu'on a beaucoup de données $M = 1000$ et 10000 (résultats en magenta et noir), on trouve aussi une coincidence lorsque la cardinalité maximale est faible ($K = 2$) et pour un nombre faible de données ($M = 100$, la première figure à gauche). Mais lorsque la cardinalité est élevée ($K = 3$ et $K = 5$) l'approche implicite est plus intéressante lorsqu'il y a peu de données ($M = 100$, résultat en bleu).
Tous ces résultats sont décrits dans la figure suivante et sont confirmés par des tests de Wilcoxon.

4.5 Conclusion

Dans ce chapitre, nous avons introduit la notion d'approche implicite pour l'apprentissage des paramètres pour trois modèles Gaussiens. Cette méthode d'estimation est similaire à celle Bayésienne, mais elle est obtenue dans un contexte naturel sans spécifier des a prioris. Cette caractéristique est intéressante pour le modèle de Heston et les réseaux Bayésiens Gaussiens conditionnels où les a prioris ne sont pas facile à établir.
L'estimation Bayésienne avec des a prioris différents des valeurs vraies peut mener à des mauvais résultats. Les estimateurs, obtenus en utilisant la méthode implicite, proposés dans ce chapitre sont alors très intéressant pour éviter de telles situations et de remplacer avantageuseument l'estimateur de maximum de vraisemblance lorsque la taille de l'échantillon est faible.

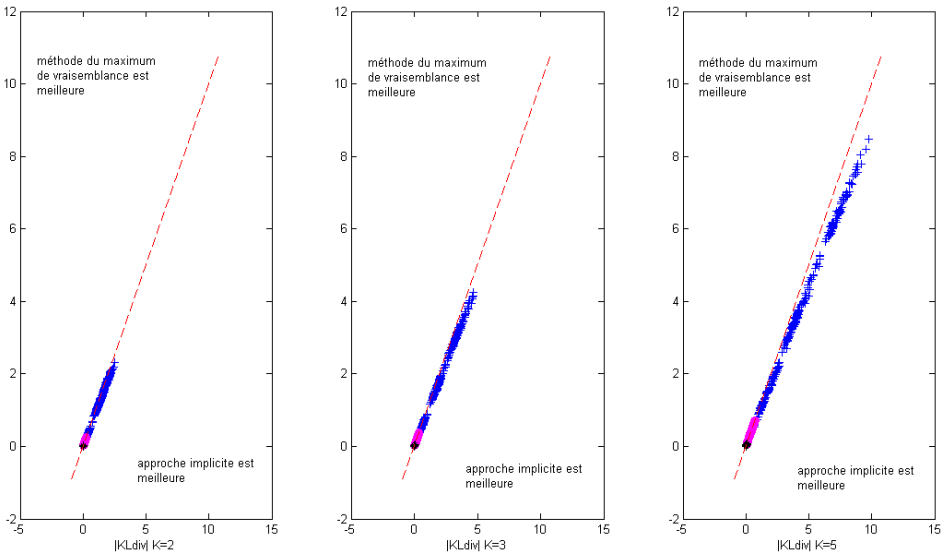

FIG. 4.2 – Comparaison de la divergence KL obtenue par l'approche Implicite contre la méthode du maximum de vraisemblance pour les mêmes ensembles de données (triangle supérieur : méthode de maximum de vraisemblance est meilleure, par rapport au triangle inférieur : approche Implicite est mieux) par rapport à la cardinalité maximale (K = 2, 3, 5) et de la taille des données ($M = 100, 1000, 10000$).

Chapitre 5

Apprentissage des paramètres de réseaux Bayésiens en mélange infini de distributions Gaussiennes

5.1 Introduction

Les modèles de mélange sont un outil très commun pour modéliser des données complexes dans de nombreux domaines. Ce succès vient de la capacité des modèles de mélanges (souvent gaussiens, mais pas toujours) à approximer des fonctions de densités associées à des variables aléatoires complexes. Dans ce chapitre, nous introduisons dans un premier temps la notion de mélange infini de lois Gaussiennes dans les réseaux Bayésiens. Nous passons ensuite à l'étude de l'apprentissage en reliant ces nouveaux modèles aux modèles linéaires non gaussiens.

5.2 Réseaux Bayésiens à mélange infini de Gaussiennes

5.2.1 Définition générale

Soit un réseau Bayésien dont les distributions conditionnelles de Y_i/X_i sont définies par

$$f_{Y_i/X_i}(y) = \int_0^\infty N(\mu_{i/X_i}(z), \sigma^2_{i/X_i}(z))(y) f(z) dz.$$

De plus, sachant $Z = z$ et X_i

$$Y_i = \beta_i(z) + \sum_{l \in \Lambda_i} \beta_{il}(z) Y_l + \sigma_{i/X_i}(z)\epsilon_i$$

avec $\epsilon_i \sim N(0,1)$, $\Lambda_i = \Big\{ l \in \{1, \ldots, n\}; Y_l \text{ est un parent de } Y_i \Big\}$.

Propriété

La moyenne μ_{i/X_i} est donnée par

$$\begin{aligned}
\mu_{i/X_i} = E(Y_i/X_i) &= \int_{\mathbb{R}} y f_{(Y_i/X_i)}(y) dy \\
&= \int_0^\infty \int_{\mathbb{R}} y N(\mu_{i/X_i}(z), \sigma^2_{i/X_i}(z))(y) f(z) dy dz \\
&= \int_0^\infty [\beta_i(z) + \sum_{l \in \Lambda_i} \beta_{il}(z) Y_l] f(z) dz \\
&= \int_0^\infty \beta_i(z) f(z) dz + \sum_{l \in \Lambda_i} Y_l \int_0^\infty \beta_{il}(z) f(z) dz.
\end{aligned}$$

Soit

$$\begin{aligned}
\widetilde{\beta}_{il} &= \int_0^\infty \beta_{il}(z) f(z) dz \\
\widetilde{\beta}_{i0} &= \int_0^\infty \beta_i(z) f(z) dz.
\end{aligned}$$

Donc

$$\mu_{i/X_i} = \widetilde{\beta}_{i0} + \sum_{l \in \Lambda_i} \widetilde{\beta}_{il} Y_l.$$

Propriété

De la même façon, la variance de Y_i sachant X_i est donnée par

$$\begin{aligned} var(Y_i/X_i) &= E(var(Y_i/X_i, Z)) + var(E(Y_i/X_i, Z)) \\ &= E(\sigma^2_{i/X_i}(Z)) + var(\mu_{i/X_i}(Z)) \\ &= \int_0^\infty \sigma^2_{i/X_i}(z) f(z) dz + var(\beta_i(z) + \sum_{l \in \Lambda_i} \beta_{il}(z) y_l) \\ &= \int_0^\infty \sigma^2_{i/X_i}(z) f(z) dz + \int_0^\infty (\beta_i(z) + \sum_{l \in \Lambda_i} \beta_{il}(z) y_l)^2 f(z) dz - \mu^2_{i/X_i}. \end{aligned}$$

5.2.2 Cas du mélange Gamma

Dans cette section et dans tout ce qui suit, nous choisissons une loi de mélange précise, la loi Gamma de paramètre (ν, λ_0) définie par

$$f(z) = \frac{\lambda_0^\nu}{\Gamma(\nu)} e^{-\lambda_0 z} z^{\nu-1} \mathbf{1}_{]0,+\infty[}.$$

On appelle ces nouveaux modèles les réseaux Bayésiens de mélange infini de Gamma Gaussienne.

Nous choisissons aussi

$$\beta_i(z) = \frac{\beta_i}{z}.$$

Théorème 5.1. *Le réseau Bayésien à mélange infini de Gamma gaussiennes est un modèle LiNGAM (cf section 1.6).*

Preuve 5.1.

$$\mu_{i/X_i} = \int_0^\infty (\frac{\beta_i}{z} + \sum_{l \in \Lambda_i} \frac{\beta_{il}}{z} y_l) \frac{\lambda_0^\nu}{\Gamma(\nu)} e^{-\lambda_0 z} z^{\nu-1} dz$$

$$\widetilde{\beta}_{il} = \int_0^\infty \beta_{il} \frac{\lambda_0^\nu}{\Gamma(\nu)} e^{-\lambda_0 z} z^{\nu-2} dz = \frac{\beta_{il} \lambda_0}{\nu - 1}$$

$$\widetilde{\beta}_{i0} = \int_0^\infty \beta_i \frac{\lambda_0^\nu}{\Gamma(\nu)} e^{-\lambda_0 z} z^{\nu-2} dz = \frac{\beta_i \lambda_0}{\nu - 1}$$

$$\mu_{i/X_i} = \widetilde{\beta}_{i0} + \sum_{l \in \Lambda_i} \widetilde{\beta}_{il} Y_l = \beta_i \frac{\lambda_0}{\nu - 1} + \sum_{l \in \Lambda_i} \beta_{il} \frac{\lambda_0}{\nu - 1} Y_l.$$

Etant donné $Y_i/Z = z \sim N(\frac{\beta_i}{z} + \frac{\sum_{l \in \Lambda_i} \beta_{il} y_l}{z}, \sigma^2_{i/X_i}(z))$ *et* $Z \sim \gamma(\nu, \lambda_0)$.
Alors la loi conditionnelle de Y_i *sachant* X_i *est définie par*

$$Y_i \sim N_\infty(\beta_i + \sum_{l \in \Lambda_i} \beta_{il} Y_l, \sigma^2_{i/X_i}, \nu, \lambda_0).$$

Donc

$$Y_i = \frac{(\beta_i + \sum_{l \in \Lambda_i} \beta_{il} Y_l)\lambda_0}{\nu - 1} + \epsilon_\infty$$

où

$$\epsilon_\infty \sim N_\infty\big(\beta_i + \sum_{l \in \Lambda_i} \beta_{il} Y_l, -\frac{(\beta_i + \sum_{l \in \Lambda_i} \beta_{il} Y_l)\lambda_0}{\nu - 1}, \sigma^2_{i/X_i}, \nu, \lambda_0\big).$$

La variance de Y_i sachant X_i est donnée par

$$var(Y_i/X_i) = \int_0^\infty \frac{\sigma^2_{i/X_i}}{z} f(z)dz + \int_0^\infty [(\frac{\beta_i}{z})^2 + (\sum_{l \in \Lambda_i} \frac{\beta_{il}}{z} y_l)^2 + 2(\frac{\beta_i}{z})$$

$$\sum_{l \in \Lambda_i} \frac{\beta_{il}}{z} Y_l] \frac{\lambda_0^\nu}{\Gamma(\nu)} e^{-\lambda_0 z} z^{\nu-1} dz - \mu^2_{i/X_i}.$$

Ceci implique que

$$var(Y_i/X_i) = \frac{\lambda_0}{\nu - 1}\sigma^2_{i/X_i} + (\beta_i + \sum_{l \in \Lambda_i} \beta_{il} Y_l)^2 (\frac{\lambda_0^2}{(\nu - 1)^2(\nu - 2)}). \tag{5.1}$$

En outre, le modèle de mélange infini est donné par

$$Y_i = \frac{\beta_i \lambda_0}{\nu - 1} + \frac{\sum_{l \in \Lambda_i} \beta_{il} \lambda_0}{\nu - 1} Y_l + \epsilon_i^\infty \tag{5.2}$$

avec ϵ_i^∞ est un mélange infini de lois gaussiennes qui est une erreur non gaussienne (Section 2.4).

5.3 Apprentissage des RBs de mélange infini de Gamma Gaussienne

Dans cette section nous nous intéressons à l'apprentissage de la structure et des paramètres des RBs de mélange infini de Gamma Gaussienne.

5.3.1 Apprentissage de la structure

Grâce à la dernière propriété (i.e. ces modèles sont des LiNGAM), nous proposons utiliser les méthodes décrites dans la section 1.6 pour apprendre la structure des RBs de mélange infini de Gamma Gaussienne. Ces méthodes donnent à la fois le graphe mais aussi les coefficients de régression pour lesquels nous proposons un estimateur spécifique dans la section suivante.

5.3.2 Apprentissage des paramètres

On a une base de données de taille N

$$\underbrace{\begin{pmatrix} Y_i^{(1)} \\ \vdots \\ Y_i^{(N)} \end{pmatrix}}_{\mathbb{Y}_i} = \frac{\lambda_0}{\nu-1} \underbrace{\begin{pmatrix} 1 & Y_1^{(1)} & \dots & Y_{q_i}^{(1)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & Y_1^{(N)} & \dots & Y_{q_i}^{(N)} \end{pmatrix}}_{\mathbb{Y}_{X_i}} \underbrace{\begin{pmatrix} \beta_i \\ \beta_{i1} \\ \vdots \\ \beta_{iq_i} \end{pmatrix}}_{\beta_i} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \epsilon_i^{\infty}$$

D'où $\mathbb{Y}_i = \frac{\lambda_0}{\nu-1}\mathbb{Y}_{X_i}\beta_i + \mathbf{1}\epsilon_i^{\infty}$.

Estimation de β_i avec la méthode des moindres carrés

On peut estimer le paramètre β_i en appliquant la méthode des moindres carrés. Cette méthode des moindres carrés consiste à minimiser la fonction risque quadratique $\varphi : \beta_i \longmapsto \|\mathbb{Y}_i - \frac{\lambda_0}{\nu-1}\mathbb{Y}_{X_i}\ \beta_i\|^2$.
La différentielle de φ par rapport à β_i

$$\begin{aligned} \frac{\partial\varphi}{\partial\beta_i} &= \frac{\partial}{\partial\beta_i}\Big(< \mathbb{Y}_i - \frac{\lambda_0}{\nu-1}\mathbb{Y}_{X_i}\beta_i, \mathbb{Y}_i - \frac{\lambda_0}{\nu-1}\mathbb{Y}_{X_i}\beta_i > \Big) \\ &= -\frac{2\lambda_0}{\nu-1} < \mathbb{Y}_i - \frac{\lambda_0}{\nu-1}\mathbb{Y}_{X_i}\beta_i, \mathbb{Y}_{X_i} > \\ &= -\frac{2\lambda_0}{\nu-1}\,{}^t\mathbb{Y}_{X_i}(\mathbb{Y}_i - \frac{\lambda_0}{\nu-1}\mathbb{Y}_{X_i}\beta_i). \end{aligned}$$

Donc

$$\widehat{\beta}_i = ({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}\mathbb{Y}_i\Big(\frac{\nu-1}{\lambda_0}\Big)$$

$$\begin{aligned} E(\widehat{\beta}_i/X_i) &= E(({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}\mathbb{Y}_i\tfrac{\nu-1}{\lambda_0}/X_i) \\ &= ({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}E(\mathbb{Y}_i/X_i)\tfrac{\nu-1}{\lambda_0} \\ &= ({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i}\beta_i \\ &= \beta_i \end{aligned}$$

D'où $\widehat{\beta}_i$ est un estimateur sans biais de β_i

$$\begin{aligned} var(\widehat{\beta}_i/X_i) &= var\Big(({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}\mathbb{Y}_i\tfrac{\nu-1}{\lambda_0}/X_i\Big) \\ &= ({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}var(\mathbb{Y}_i/X_i)\tfrac{(\nu-1)^2}{(\lambda_0)^2}\mathbb{Y}_{X_i}({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1} \\ &= ({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1}\ {}^t\mathbb{Y}_{X_i}var(Y_i/X_i)\mathbb{Y}_{X_i}({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1} \\ &= var(Y_i/X_i)({}^t\mathbb{Y}_{X_i}\mathbb{Y}_{X_i})^{-1} \end{aligned}$$

Estimation de β_i par approche LiNGAM

L'algorithme d'apprentissage LiNGAM retourne les coefficients de régression et donc

$$\beta_i = \frac{a_i(\nu - 1)}{\lambda_0 \; a_{X_i}}$$

Estimation de la variance conditionnelle

Rappelons que $var(Y_1, \ldots, Y_n/Z = z) = \frac{\Sigma}{z}$ et $E(Y_1, \ldots, Y_n/Z = z) = \frac{a}{z}$, avec $a = {}^t(a_1, \ldots, a_n)$. Donc $\Sigma_\infty = var(Y_1, \ldots, Y_n) = \frac{\lambda_0 \Sigma}{\nu - 1} + \frac{A \otimes A \lambda_0^2}{(\nu-1)^2(\nu-2)}$
où

$$A = (\beta_i + \sum_{l \in \Lambda_i} \beta_{il} Y_l)_{1 \leq i \leq n} \in \mathbb{R}^n \qquad and \qquad A \otimes A = A \; {}^t A$$

en utilisant la méthode des moments, on obtient

$$\widehat{\Sigma}_\infty = S_N^2 = \frac{1}{N} \sum_{h=1}^{N} (Y^{(h)} - \overline{Y}) \; {}^t(Y^{(h)} - \overline{Y})$$

où

$$Y^{(h)} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(N)} \end{pmatrix}$$

avec $Y^{(1)}, \ldots, Y^{(N)}$ sont des copies indépendantes de Y.
Et par la suite

$$\widehat{\Sigma} = \frac{\nu - 1}{\lambda_0} (S_N^2 - \frac{A \otimes A \lambda_0^2}{(\nu - 1)^2 (\nu - 2)})$$

d'où l'estimation de $\widehat{\sigma}^2_{i/X_i}$ est donnée par

$$\widehat{\sigma}^2_{i/X_i} = \widehat{\Sigma}_i - \widehat{\Sigma}_{iX_i} \widehat{\Sigma}^{-1t}_{X_i} \widehat{\Sigma}_{iX_i}$$

où $\widehat{\Sigma}_i$ est l'estimation de la variance inconditionnelle de Y_i, $\widehat{\Sigma}_{iX_i}$ est la matrice de covariance estimée entre Y_i et les variables X_i, et $\widehat{\Sigma}_{X_i}$ est la matrice de covariance estimée X_i.

5.4 Travaux apparentés

Dans un modèle de mélange il n'est pas nécessaire de limiter le nombre de composantes d'être fini.

Rasmussen [60] a présenté un modèle de mélange infini de gaussienne qui évite nettement le problème difficile de trouver le "bon" nombre de composantes du mélange. L'inférence dans le modèle est faite en utilisant l'échantillonnage de Gibbs.
Un modèle de mélange de processus de Dirichlet peut être construit comme une limite d'un modèle de mélange paramétrique. Rasmussen *et al* [24] établient la formulation du modèle de mélange gaussien hiérarchique et ensuite tendre la limite comme le nombre de composantes du mélange vers l'infini pour obtenir le modèle mélange de processus de Dirichlet.
Le modèle de mélange de processus de Dirichlet étend le modèle de mélange traditionnel afin d'avoir un nombre infini de composants. Les poids de mélange et les vecteurs de paramètres des composants sont traités comme des quantités aléatoires.

5.5 Expérimentation

5.5.1 Protocole expérimental

Nous avons effectué des simulations dans plusieurs contextes.
Dans le premier contexte, la structure et les paramètres de régression sont appris par LiNGAM (LiNGAM 1). Dans le deuxième contexte, la structure est appris par LiNGAM puis nous ajoutons l'hypothèse de mélange infini de Gamma Gaussienne pour aboutir à l'apprentissage des paramètres (LiNGAM 2). Dans le troisième contexte, nous gardons le graphe théorique puis nous faisons l'apprentissage des paramètres avec le mélange infini de Gaussienne (notre méthode).
Dans ces contextes, nous avons contrôlé plusieurs paramètres tels que le nombre n de variable n=(10, 30, 50) et la taille M de l'ensemble des données générées M=(100, 1000, 10000). La cardinalité maximale K de nos variables est également controlée par les modèles LiNGAM ($K = 1$).
Chaque génération de données dans ces conditions est itéré 10×10 fois, avec 10 DAG générés aléatoirement et 10 valeurs aléatoires des paramètres pour chacun de ces DAG.
Dans cette partie expérimentale nous avons implémenté nos différents mo-

dèles et algorithmes en utilisant Matlab et plus précisément la Bayes Net Toolbox.

5.5.2 Résultats et Interprétations

Nous comparons l'erreur quadratique pour l'estimation des paramètres de régression dans les trois contextes à l'égard du nombre de variables n et le nombre de données M.
Nous nous intéressons à comparer en premier lieu notre méthode avec LiNGAM 2. LiNGAM 2 est moins bon que notre méthode puisque le LiNGAM 2 apprend la structure et les paramètres et nous on part avec la structure théorique donc on est plus avantagé. Nos résultats sont décrits dans la figure 5.1. On constate que notre méthode est meilleure pour toutes les valeurs de n et M.
En comparant LiNGAM 1 avec LiNGAM 2 (figure 5.2) on constate une concordance entre les deux méthodes pour $M = 1000$ (résultats en magenta) et $M = 10000$ (résultats en noir), mais lorsque le nombre de données est faible $M = 100$ (résultats en bleu) le LiNGAM 2 donne des résultats mieux que le LiNGAM 1.

5.6 Conclusion

Nous avons introduit dans ce chapitre une nouvelle classe de réseaux bayésiens, dont les distributions conditionnelles sont des mélanges infinis de gaussiennes, puis nous avons étudié l'apprentissage de tels modèles. De plus, nous avons montré que ces modèles sont des LiNGAM. Ensuite nous avons présenté une partie expérimentale pour valider notre méthode d'estimation.

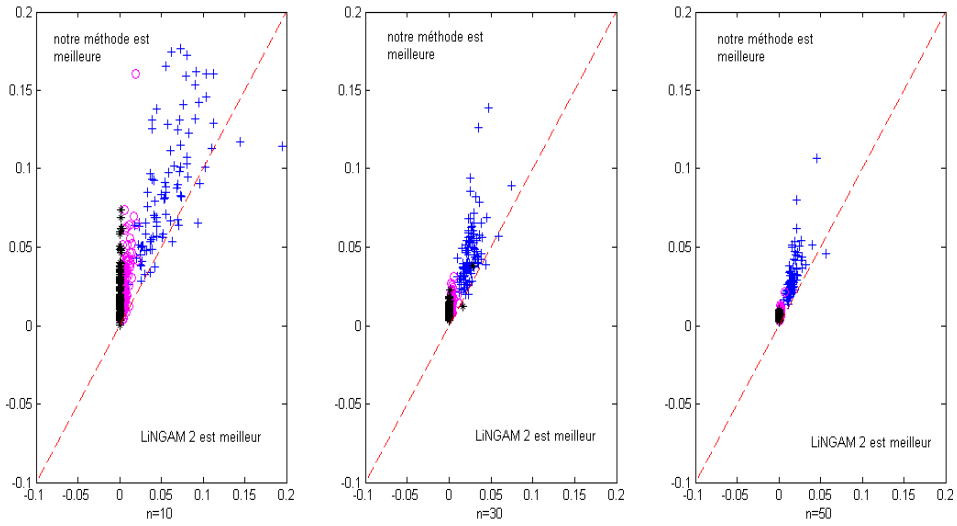

FIG. 5.1 – Comparaison de l'erreur quadratique obtenue par notre méthode contre LiNGAM 2 pour les mêmes ensembles de données (triangle supérieur : notre méthode est meilleure, par rapport au triangle inférieur : LiNGAM 2 est meilleur) par rapport au nombre de variables (n = 10, 30, 50) pour $M = 100, 1000, 10000$.

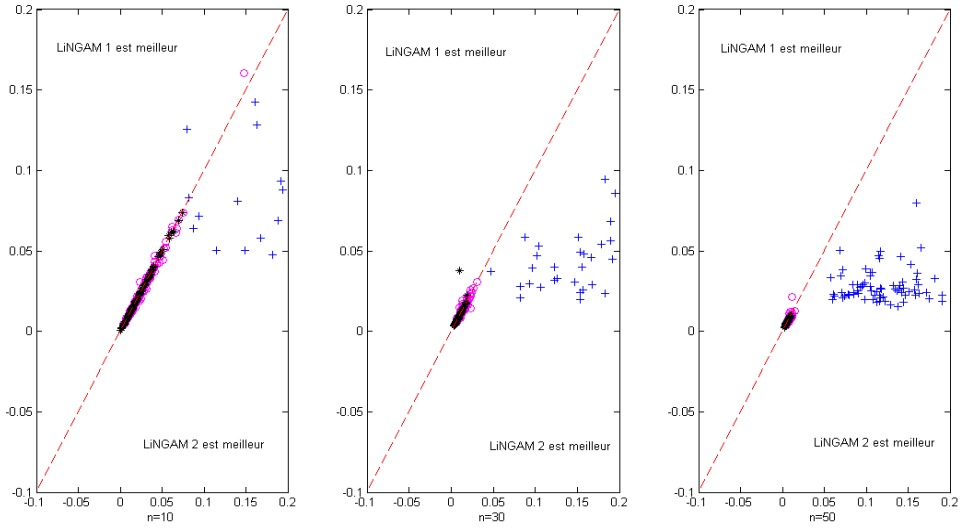

FIG. 5.2 – Comparaison de l'erreur quadratique obtenue par LiNGAM 1 contre LiNGAM 2 pour les mêmes ensembles de données (triangle supérieur : LiNGAM 1 est meilleur, par rapport au triangle inférieur : LiNGAM 2 est meilleur) par rapport au nombre de variables (n = 10, 30, 50) pour $M = 100, 1000, 10000$.

Conclusion et perspectives

Cette thèse porte sur des nouvelles formes de paramétrisation de réseaux Bayésiens, avec trois contributions.
La première contribution est basée sur les réseaux Bayésiens exponentiels discrets. Nous avons résolu les problèmes d'apprentissage des paramètres et de la structure. Nous avons développé une famille des a prioris qui généralise les a prioris Dirichlet utilisés pour la loi multinomiale, et par la suite nous avons obtenu une fonction score qui étend le score Bayésien Dirichlet. Une étude de simulation a déterminé de façon empirique dans quels contextes les réseaux Bayésiens exponentiels discrets peuvent être une bonne alternative par rapport aux réseaux Bayésiens usuels. Les expériences décrites ici nous montrent que les RBeds Poisson peuvent être une bonne alternative aux RBs usuels pour l'estimation de la densité lorsque la taille de l'échantillon est faible ou lorsque la cardinalité maximale des variables est élevée, en raison de la réduction du nombre de paramètres utilisés par les RBeds.
La seconde contribution concerne l'application de la méthode implicite dans plusieurs modèles. Nous appliquons d'abord cette dernière approche pour l'apprentissage des paramètres dans un modèle à volatilité stochastique. Ensuite, en reprenant les travaux de Murphy [55] sur l'estimation des réseaux Bayésiens Gaussiens conditionnels, nous avons complété les approches classiques en proposant les estimateurs Bayésiens de type espérance a posteriori puis les estimateurs implicites des mêmes paramètres et nous avons effectué une étude comparative de ces estimateurs. Nos résultats sont illustrés par une étude de simulation.
La dernière contribution concerne l'utilisation de mélanges infini de Gaussiennes. Nous avons proposé une paramétrisation non Gaussienne qui com-

pléte les modèles LiNGAM (Linear, Non-Gaussian, Acyclic causal Models). Nous avons aussi proposé des estimateurs des paramètres de ces mélanges inifinis.

Perspectives

Les expériences décrites pour les RBeds Poisson pourraient être étendues à d'autres distributions de familles exponentielles naturelles discrètes telle que la famille de lois binomiale négative, ou des lois absolument continues. Pour chaque loi de probabilité, nous avons besoin de proposer une meilleure façon de traiter les paramètres a priori tels que t_{ij} et m_{ij} de la loi de Poisson afin d'obtenir la fonction score vérifiant la propriété d'équivalence de Markov (comme la fonction score BDe pour les RBs discrets usuels).

L'apprentissage de la strucure d'un réseau Bayésien à partir de la base de données est un des défis les plus existants. Nous proposons de nous baser sur l'approche Implicite pour étudier l'apprentissage de structure pour les réseaux Bayésiens Gaussiens conditionnels.

Un domaine à étudier est celui des modèles LinGAM (Linear, Non-Gaussian, Acyclic causal Models) récemment proposés par [64] pour lesquels il n'existe à notre connaissance aucun algorithme d'apprentissage de structure à base de score. Ces modèles, extensions des modèles SEM (Structural Equation Model) possèdent des propriétés théoriques très intéressantes puisqu'il est possible d'identifier entièrement la structure graphique à partir de données d'observation, ce qui n'est pas le cas dans le cas continu standard des modèles gaussiens.

Pour les FENs ou les mélanges infinis, nous avons proposé de nouvelles formes paramétriques de RB, pour lesquels se pose aussi le problème de l'inférence probabiliste. Cela ne devrait pas trop poser de problème pour les FENs. Les algorithmes d'inférence probabiliste doivent également être étendus pour ces lois de probabilités, ce qui semble ne pas poser trop de difficultés pour une distribution d'une famille exponentielle naturelle comme il est montré dans [43] pour les RBs hybrides avec distribution Gaussienne conditionnelle.

Pour les mélanges infinis, des méthodes approchées par échantillonnage ou variationnelles pourraient être aussi proposées.

Bibliographie

[1] L. Andersen. Efficient simulation of the heston stochastic volatility model. *Banc of America*, pages 1–38, 2006.

[2] O. Barndorff-Nielsen. *Information and Exponential families in Statistical Theory*. John Wiley, 1978.

[3] M. Beal and Z. Ghahramani. The variational bayesian em algorithm for incomplete data : with application to scoring graphical model structures. *Bayesian Statistics*, 7 :453–464, 2003.

[4] H. Ben Hassen, A. Masmoudi, and A. Rebai. Causal inference in biomolecular pathways using a bayesian network approach and an implicit method. *Journal of Theoretical Biology*, 253(4) :717 – 724, 2008.

[5] H. Ben Hassen, A. Masmoudi, and A. Rebai. Inference in signal transduction pathways using em algorithm and an implicit algorithm : Incomplete data case. *Journal of Computational Biology*, 16(9) :1227–1240, 2009.

[6] J. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley, New York, 1994.

[7] F. Black and M.S. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81 :637–654, 1973.

[8] C. Bruni and G. Koch. Identifiability of continuous mixtures of unknown gaussian distributions. *The Annals of Probability*, 13(4) :1341–1357, 1985.

[9] D. Chickering. Learning bayesian networks is np-complete. In *Proceedings of AI and Statistics*, pages 121–130, 1995.

[10] D. Chickering, D. Geiger, and D. Heckerman. Learning bayesian networks : Search methods and experimental results. In *Proceedings of Fifth Conference on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[11] D. Chickering and D. Heckerman. Efficient Approximation for the Marginal Likelihood of Incomplete Data given a Bayesian Network. In *UAI'96*, pages 158–168. Morgan Kaufmann, 1996.

[12] D.M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3 :507–554, November 2003.

[13] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3) :462–467, 1968.

[14] G. Consonni and P. Veronese. Conjugate priors for exponential families having quadratic variance functions. *Amer. Statist. Assoc*, 87 :1123–1127, 1992.

[15] G. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9 :309–347, 1992.

[16] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[17] J.C. Cox, J.E. Ingersoll, and S.A. Ross. A theory of a term structure of interest rates. *Econometrica*, 53 :385–407, 1985.

[18] R. Daly, Q. Shen, and S. Aitken. Learning bayesian networks : approaches and issues. *The knowledge Engineering Review*, 26(2) :99–157, 2011.

[19] G. Deelstra and F. Delbaen. Convergence of discretized stochastic (interest rate) processes with stochastic drift term. *Appl. Stochastic Models Data Anal*, 14 :77–84, 1998.

[20] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *Ann. Statist*, 7 :269–281, 1979.

[21] R.A. Fisher. The fiducial argument in statistical inference. *Ann. Eugen*, 6 :391 – 398, 1935.

[22] D. Geiger and D. Heckerman. Learning gaussian networks. Technical report, Redmond, WA, 1994.

[23] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families : graphical models and model selection. *Annals of Statistics*, 29 :505–529, 2001.

[24] D. Görür and C. E. Rasmussen. Dirichlet process gaussian mixture models : Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4) :653–664, 2010.

[25] A. Graja, A. Jarraya, and A. Masmoudi. Implicit estimation for stochastic volatility model. *Communications in Statistics Theory and Methods*, 2012.

[26] A. Hassairi, A. Masmoudi, and C Kokonendji. Implicit distributions and estimation. *Communications in Statistics - Theory and Methods*, 34(2) :245 – 252, 2005.

[27] D. Heckerman. A tutorial on learning with bayesian network. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. Kluwer Academic Publishers, Boston, 1998.

[28] D. Heckerman and D. Geiger. Learning bayesian networks : A unification for discrete and gaussian domains. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 274–284, San Francisco, CA, 1995. Morgan Kaufmann Publishers.

[29] D. Heckerman, D. Geiger, and M. Chickering. Learning Bayesian networks : The combination of knowledge and statistical data. In Ramon Lopez de Mantaras and David Poole, editors, *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 293–301, San Francisco, CA, USA, July 1994. Morgan Kaufmann Publishers.

[30] S. Heston. A closed form solution of options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6 :327–343, 1993.

[31] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, E. Reiman, and the AlzheimerŠs Disease Neuroimaging Initiative. A sparse structure learning algorithm for gaussian bayesian network identification from high-

dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6) :1328–1342, 2013.

[32] A. Hyvarinen, J. Karhumen, and E. Oja. Independent component analysis. *Wiley, New York*, 2001.

[33] A. Jarraya, P. Leray, and A. Masmoudi. Discrete exponential bayesian networks : an extension of bayesian networks to discrete natural exponential families. In *23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI'2011)*, pages 205–208, Boca Raton, Florida, USA, 2011.

[34] A. Jarraya, P. Leray, and A. Masmoudi. Discrete exponential bayesian networks structure learning for density estimation. In *Eighth International Conference on Intelligent Computing (ICIC 2012)*, pages 146–151, 2012.

[35] A. Jarraya, P. Leray, and A. Masmoudi. A new "implicit" parameter estimation for conditional gaussian bayesian networks. In *The 10th International FLINS Conference on Uncertainty Modeling in Knowledge Engineering and Decision Making (FLINS 2012)*, Istanbul, Turkey, 2012.

[36] A. Jarraya, P. Leray, and A. Masmoudi. Discrete exponential bayesian networks : definition, learning and application for density estimation. *Neurocomputing*, 2013.

[37] A. Jarraya, P. Leray, and A. Masmoudi. Implicit parameter estimation for conditional gaussian bayesian networks. *International Journal of Computational Intelligence Systems*, 6 :6–17, 2013.

[38] F.V. Jensen. *An introduction to Bayesian Networks*. Taylor and Francis, London, United Kingdom, 1996.

[39] D. Koller and N. Friedman. *Probabilistic Graphical Models : Principles and Techniques*. MIT Press (MA), 2009.

[40] S. Lauritzen. Propagation of probabilistics, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87 :1098–1108, 1992.

[41] S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[42] S. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17 :31–57, 1989.

[43] S.L. Lauritzen and F. Jensen. Stable local computation with conditional Gaussian distributions. *Statistics and Computing*, 11(2) :191–203, April 2001.

[44] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2) :157–224, 1988.

[45] P. Leray. Réseaux bayésiens : Apprentissage et diagnostic de systemes complexes, Habilitation à Diriger les Recherches, Université de Rouen, France, 2006.

[46] P. Leray and O. Francois. BNT structure learning package : Documentation and experiments. Technical report, Laboratoire PSI, 2004.

[47] P. Leray and O. Francois. Bayesian network structural learning and incomplete data. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005)*, pages 33–40, Espoo, Finland, 2005.

[48] G. Letac. Lectures on natural exponential families and their variance functions. *Monograph. Math*, 5, 1992.

[49] G. J. Mclachlan and D. Peel. Finite mixture models. *Wiley, New York*, 2000.

[50] S. Monti and G. F. Cooper. Learning hybrid bayesian network from data. *In Michael I. Jordan, editor, Learning in Graphical Models. Kluwer Academic Publishers, Boston*, pages 521–540, 1998.

[51] C. N. Morris. Natural exponential families with quadratic variance functions. *Ann. Statist*, 10 :65–80, 1982.

[52] N. Mukhopadhyay. Bayesian inference some comments on hassairi et al.s implicit distributions and estimation commun. *Statist.Theor. Meth*, 35 :293–207, 2006.

[53] K. Murphy. Inference and learning in hybrid bayesian networks. Technical report, 1998.

[54] K. Murphy. The Bayes Net Toolbox for Matlab. In *Computing Science and Statistics : Proceedings of Interface*, volume 33, 2001.

[55] K. Murphy. Fitting a conditional gaussian distribution. Technical report, UC Berkeley, 2003.

[56] P. Naïm, O. Pourret, and B. Marcot. Dirichlet process gaussian mixture models : Choice of the base distribution. *Bayesian Networks : A Practical Guide to Applications, Wiley*, 2008.

[57] P. Naïm, P-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux bayésiens*. Eyrolles, Paris, 3 edition, 2007.

[58] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.

[59] J. Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[60] C. E. Rasmussen. The infinite gaussian mixture model. *in Advances in Neural Information Processing Systems*, 12 :554–560, 1999.

[61] C.P. Robert. *The Bayesian Choice : a decision-theoretic motivation*. Springer, New York, 1994.

[62] R. W. Robinson. Counting unlabeled acyclic digraphs. In C. H. C. Little, editor, *Combinatorial Mathematics V*, volume 622 of *Lecture Notes in Mathematics*, pages 28–43, Berlin, 1977. Springer.

[63] M. Schmidt, A.N Mizil, and K. Murphy. Learning graphical model structure using l1-regularization paths. *Association for the Advancement of Artificial Intelligence*, pages 1278–1283, 2007.

[64] S. Shimizu, A. Hyvarinen, Y. Kano, and P.O. Hoyer. Discovery of non-gaussian linear causal models using ica. *In Proc. the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 526–533, 2005.

[65] S. Shimizu, A. Hyvarinen, Y. Kawahara, and T. Washio. A direct method for estimating a causal ordering in a linear non-gaussian acyclic model. *in Proc. 25th Conf. on Uncertainty in Artificial Intelligence (UAI2009), (Montreal, Canada)*, pages 506–513, 2009.

[66] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, editors, *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 220–227, San Francisco, 1991. Morgan Kaufmann.

[67] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1 :1–305, 2008.

Annexe A

Dans cette Annexe on introduit les différentes densités utilisées dans cette thèse.

1. Loi Normale "$N(\mu, \sigma^2)$"

Soit X une variable aléatoire de loi normale $N(\mu, \sigma^2)$ (On note $X \sim N(\mu, \sigma^2)$). La densité de X est donnée par

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{(x-\mu)^2}{2\sigma^2} \Big\}$$

où $\mu \in \mathbb{R}$ et $\sigma > 0$. Les premiers moments de la loi normale $N(\mu, \sigma^2)$ sont

$$\begin{cases} E(X) &= \mu \\ V(X) &= \sigma^2. \end{cases}$$

2. Loi Gamma "$\gamma(\alpha, \lambda)$"

Soit X une variable aléatoire de loi gamma $\gamma(\alpha, \lambda)$ de paramètres $\alpha > 0$ et $\lambda > 0$. La densité de X est donnée par

$$p(x) = \frac{\lambda^\alpha x^{\alpha-1} \exp\{-\lambda x\}}{\Gamma(\alpha)} \mathbf{1}_{]0,+\infty[}(x)$$

où $\Gamma(\alpha) = \displaystyle\int_0^{+\infty} t^{\alpha-1} e^{-t} dt$.

Les premiers moments de la loi $\gamma(\alpha, \lambda)$ sont

$$\begin{cases} E(X) &= \dfrac{\alpha}{\lambda} \\ V(X) &= \dfrac{\alpha}{\lambda^2}. \end{cases}$$

3. Loi Inverse Gamma "$IG(a, b)$"

Soit X une variable aléatoire de loi Inverse Gamma $IG(a, b)$ de paramètres $a > 0$ et $b > 0$. La densité de X est donnée par

$$p(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\{\frac{-b}{x}\} \mathbf{1}_{]0,+\infty[}(x)$$

Les premiers moments de la loi $IG(a,b)$ sont

$$\begin{cases} E(X) & = \dfrac{b}{a-1} \quad (a>1) \\ V(X) & = \dfrac{b^2}{(a-1)^2(a-2)} \quad (a>2). \end{cases}$$

Notons que si $X \sim \gamma(\lambda,\nu)$ alors $\frac{1}{X} \sim IG(\lambda, \frac{1}{\nu})$.

4. Loi Multinomiale

La variable aléatoire $(X_1,\ldots,X_k)$ suit la loi Multinomiale de dénominateur n et probabilités $(p_1,\ldots,p_k)$ si sa fonction de probabilité est

$$P(x_1,\ldots,x_k) = \frac{n!}{x_1!\times\ldots\times x_k!} p_1^{x_1} p_2^{x_2}\ldots p_k^{x_k}, \quad x_1,\ldots,x_k \in \{0,\ldots,n\}$$

où $n \in \mathbb{N}$ et $p_1,\ldots,p_k \in [0,1]$ avec $p_1+\ldots+p_k=1$

On a les moments suivants pour $i \neq j = 1,\ldots,k$

$$\begin{cases} E(X_i) & = np_i \\ V(X_i) & = np_i(1-p_i) \\ Cov(X_i,X_j) & = -np_ip_j \end{cases}$$

5. Densité de Dirichlet

La loi de Dirichlet d'ordre $k \geq 2$ de paramètres $\alpha_1,\ldots,\alpha_k > 0$ possède pour densité de probabilité

$$P(x_1,\ldots,x_k) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i-1}$$

pour tous les $x_1,\ldots,x_k > 0$ vérifiant $x_1+\ldots+x_{k-1} < 1$ où x_k est une abréviation pour $1-x_1-\ldots-x_{k-1}$. La densité est nulleen dehors de ce simplexe ouvert de dimension $(k-1)$.

Les moments principaux sont, pour $i \neq j$

$$\begin{cases} E(x_i) & = \frac{\alpha_i}{\alpha_1+\ldots+\alpha_k} \\ V(X_i) & = \frac{\alpha_i(\alpha_1+\ldots+\alpha_k-\alpha_i)}{(\alpha_1+\ldots+\alpha_k)^2(\alpha_1+\ldots+\alpha_k+1)} \\ Cov(X_i,X_j) & = -\frac{\alpha_i\alpha_j}{(\alpha_1+\ldots+\alpha_k)^2(\alpha_1+\ldots+\alpha_k+1)} \end{cases}$$

6. Loi Wishart "$W_p(n,V)$"

Soit X une variable aléatoire de loi Wishart $W_p(n,V)$ de paramètres $n > p-1$

et $V > 0$.

La densité de X est donnée par

$$p(x) = \frac{1}{2^{\frac{np}{2}}|V|^{\frac{n}{2}}\Gamma(\frac{n}{2})}|X|^{\frac{n-p-1}{2}}e^{-\frac{1}{2}tr(V^{-1}X)}$$

7. Loi Inverse Wisnart "$W^{-1}(\nu, \Psi)$"

Soit X une variable aléatoire de loi Inverse Wishart $W^{-1}(\nu, \Psi)$ de paramètres $\nu > p - 1$ et $\Psi > 0$.

La densité de X est donnée par

$$\frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma(\frac{\nu}{2})}|X|^{-\frac{\nu+p+1}{2}}e^{-\frac{1}{2}tr(\Psi X^{-1})}$$

Annexe B

L'approche Bayésienne

Dans l'approche classique de la statistique, les paramètres à estimer, bien qu'inconnus sont supposés fixes. On procède souvent par la méthode du maximum de vraisemblance (ML) pour les innovations gaussiennes. L'idée dans ce cas est de déterminer les paramètres qui cadrent le mieux avec des données observées, étant donnée la fonction de vraisemblance. On obtient alors des valeurs pour différents paramètres et on peut directement procéder aux prévisions sans tenir compte de l'incertitude qui existe dans les paramètres ainsi estimés. Dans une optique différente, l'approche Bayésienne cherche à formaliser autrement l'incertitude relative aux paramètres inconnus d'un modèle. Le statisticien part d'une loi a priori des paramètres et obtient une loi a posteriori plus précise à la lumière des données. Les prévisions sont alors basées sur cette loi a posteriori, prenant ainsi en compte l'incertitude qui plane sur les paramètres. Cette approche nous semble assez naturelle dans la mesure où elle reflète l'idée d'apprentissage constatée dans la vie quotidienne. En effet, le statisticien part avec une idée d'un phénomène et révise ses anticipations au fur et à mesure qu'il observe les tendances du phénomène. Ce qui différencie une situation statistique d'une situation Bayésienne est qu'on tient compte d'une information a priori sur un paramètre inconnu θ.
Plus précisément, soit X une variable aléatoire suivant une loi de probabilité notée P_θ ; $\theta \in \Theta$. On munit l'ensemble des paramètres Θ d'une tribu et d'une probabilité qu'on appelle loi a priori de θ reflétant les croyances sur les valeurs plausibles de θ qui devient alors une variable aléatoire qui suit une loi η dite loi a priori. Alors, P_θ devient la loi conditionnelle de l'observation

X sachant θ. On note $P_\theta = P(X/\theta)$.
La loi du couple (X, θ) s'exprime à l'aide de P_θ et de η :

$$P(X, \theta) = P_\theta \otimes \eta.$$

Si on se place dans un modèle dominé ($P_\theta << \mu$), on adopte les notations suivantes

$$p(x/\theta) = \frac{dP_\theta}{d\mu}(x).$$

On suppose que η est dominée par une mesure υ et on pose

$$\pi(\theta) = \frac{d\eta}{d\upsilon}(\theta).$$

Ainsi, la densité du couple (X, θ) par rapport à $\mu \otimes \upsilon$ est

$$p(x, \theta) = p(x/\theta)\pi(\theta).$$

On peut donc calculer $P(\theta/X)$ qu'on appelle loi a posteriori de θ reflétant les croyances actualisées après l'observation des données X. Avec les notations qu'on vient d'introduire, la loi a posteriori admet une densité par rapport à υ qui est égale à

$$p(\theta/x) = \frac{p(x/\theta)\pi(\theta)}{\displaystyle\int_\Theta p(x/\theta)\pi(\theta)d\upsilon(\theta)}. \tag{5.3}$$

Lorsqu'on s'intéresse à la loi de θ sachant X, le dénominateur de l'équation (5.3) sera une constante vu que la loi est conditionnelle à X.
Cette dernière relation est fondamentale pour l'analyse Bayésienne car elle nous permet d'avoir la loi a posteriori des paramètres à un facteur multiplicatif près (i.e ; $p(\theta/x) \propto p(x/\theta)\pi(\theta)$).
Bien que la distribution a posteriori de θ soit importante, il est peut être nécessaire de résumer la distribution a posteriori sous une valeur de θ "optimale", comme dans l'approche du maximum de vraisemblance. Cette valeur "optimale" n'est pas unique puisqu'elle est liée à une fonction perte qui est due au choix du meilleur estimateur de θ.

Définition 5.1. *Soit T une statistique. Une application définie par*

$$\begin{array}{rccc} \rho : & \Omega \times \Theta & \longrightarrow & \mathbb{R}_+ \\ & (w, \theta) & \longmapsto & \rho(T(w), \theta) \end{array}$$

est dite fonction perte. Elle mesure le coût d'estimation de θ par $T(w)$.

Définition 5.2. *Soit ρ une fonction perte ; l'application définie par*

$$\begin{array}{rccc} R(T,.): & \Theta & \longrightarrow & \mathbb{R}_+ \\ & \theta & \longmapsto & E_\theta(\rho(T,\theta)) \end{array}$$

est appelée fonction risque.
Le risque Bayésien est donc défini comme

$$R(T) = \int_\Theta R(T,\theta)\eta(d\theta).$$

Définition 5.3. *On appelle estimateur Bayésien associé à une loi a priori et à une fonction perte dans le modèle statistique $\{P_\theta, \theta \in \Theta\}$, l'estimateur $\widehat{\theta}$ (s'il existe) qui minimise le risque Bayésien.*

Proposition 5.1. *Si $\rho(T,\theta) = (T-\theta)^2$ c'est-à-dire la perte quadratique, alors l'estimateur Bayésien $\widehat{\theta}$ de θ est égal à la moyenne de la loi a posteriori de θ sachant X (i.e ; $\widehat{\theta} = E(\theta/X)$).*

Exemple Soit $(X_1, \ldots, X_n)$ un n-échantillon aléatoire de loi normale $N(\theta, 1)$ de moyenne θ et de variance 1. On suppose que la loi a priori de θ est la loi Normale $N(m, \sigma^2)$ de moyenne m et de variance σ^2.
Ainsi, la loi a posteriori de θ sachant $X_1 = x_1, \ldots, X_n = x_n$ est aussi une loi Normale

$$N\Big(\frac{m + \sigma^2 \sum_{i=1}^{n} x_i}{1 + n\sigma^2}, \frac{\sigma^2}{n\sigma^2 + 1}\Big).$$

En considérant la perte quadratique, l'estimateur Bayésien $\widehat{\theta}$ de θ est alors

$$\widehat{\theta} = E(\theta/X_1, \ldots, X_n) = \frac{m + \sigma^2 \sum_{i=1}^{n} X_i}{1 + n\sigma^2}.$$

Définition 5.4. *Soit $(P_\theta)_{\theta\in\Theta}$ un modèle statistique et soit $R = (R_s)_{s\in S}$ une famille de loi sur Θ. On dit que cette famille est conjuguée à $(P_\theta)_{\theta\in\Theta}$ si à chaque fois que la loi a priori est un élément de R, la loi a posteriori est aussi dans R.*

Au niveau des concepts, l'approche Bayésienne [61] est donc simple, naturelle et flexible. Elle consiste à aboutir, à partir d'une loi a priori assez imprécise

des paramètres, à une loi a posteriori plus précise à la lumière des données. Le problème pratique est de spécifier l'a priori $\eta(d\theta)$ en tenant compte des informations dont on dispose, de caractériser correctement comme dans l'approche du maximum de vraisemblance, la vraissemblance $P(X/\theta)$ et enfin de calculer l'a posteriori $P(\theta/X)$.

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Aida JARRAYA SIALA

Nouvelles paramétrisations de réseaux Bayésiens et leur estimation implicite

Famille exponentielle naturelle et mélange infini de Gaussiennes

New parametrisation of Bayesian networks and their implicit estimation

Natural exponential family and Gaussian infinite mixture

Résumé
L'apprentissage d'un réseau Bayésien consiste à estimer le graphe (la structure) et les paramètres des distributions de probabilités conditionnelles associées à ce graphe. Les algorithmes d'apprentissage de réseaux Bayésiens utilisent en pratique une approche Bayésienne classique d'estimation a posteriori dont les paramètres sont souvent déterminés par un expert ou définis de manière uniforme Le coeur de cette thèse concerne l'application aux réseaux Bayésiens de plusieurs avancées dans le domaine des Statistiques comme l'estimation implicite, les familles exponentielles naturelles ou les mélanges infinis de lois Gaussiennes dans le but de (1) proposer de nouvelles formes paramétriques, (2) estimer des paramètres de tels modèles et (3) apprendre leur structure.

Mots clés
Réseau bayésien, Estimation implicite, Famille exponentielle, Mélange infini de gausiennes

Abstract
Learning a Bayesian network consists in estimating the graph (structure) and the parameters of conditional probability distributions associated with this graph. Bayesian networks learning algorithms rely on classical Bayesian estimation approach whose a priori parameters are often determined by an expert or defined uniformly The core of this work concerns the application of several advances in the field of statistics as implicit estimation, Natural exponential families or infinite mixtures of Gaussian in order to (1) provide new parametric forms for Bayesian networks, (2) estimate the parameters of such models and (3) learn their structure.

Key Words
Bayesian network, Implicit estimation, Exponential family, Gaussian infinite mixtures